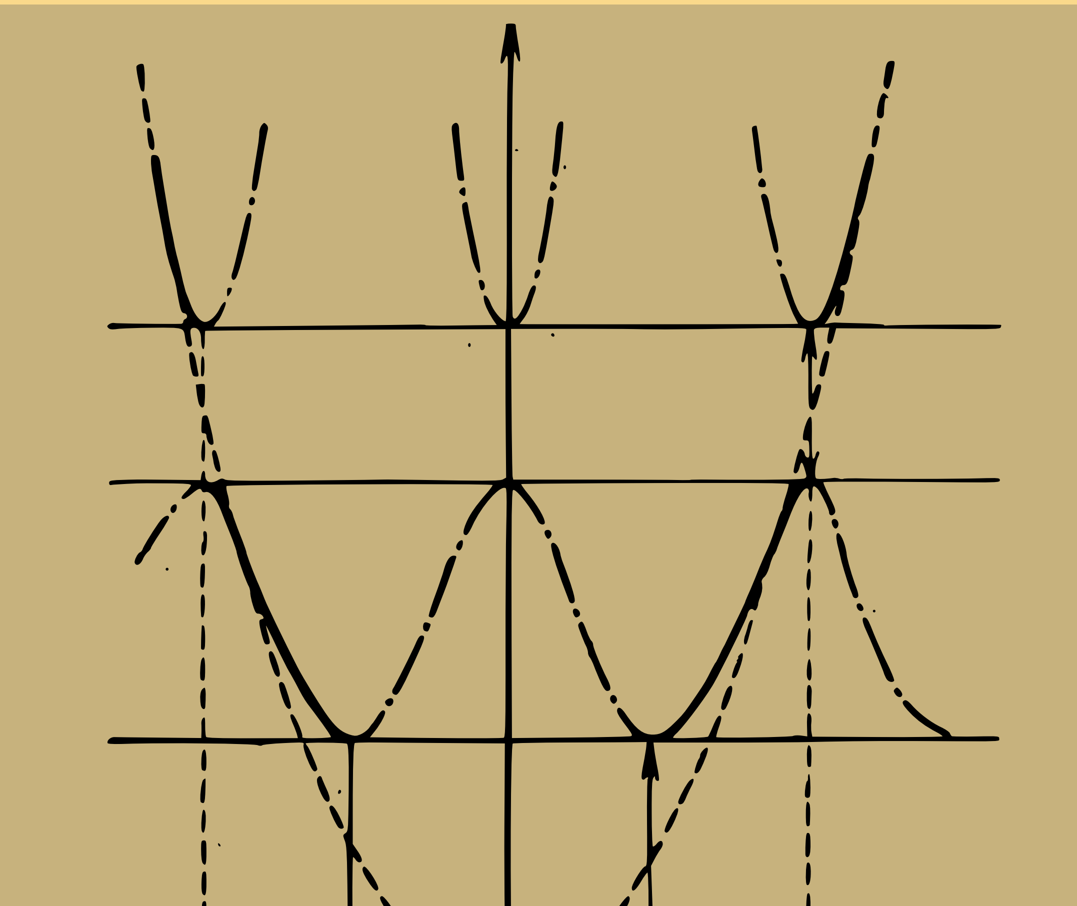
P.S. Kireev

Semiconductor Physics



Mir Publishers

Semiconductor Physics

P. S. Kireev

105

P. S. KIREEV

SEMICONDUCTOR PHYSICS

Translated from the Russian
by
Mark Samokhvalov

MIR PUBLISHERS
MOSCOW

First published 1975
Second edition 1978

На английском языке

OTHER MIR PUBLISHERS TITLES

Introduction to Physics

A. I. KITAIGORODSKY, D. Sc.

This book, written by a leading Soviet physicist, is an educational aid for physics students at institutions of higher learning as well as for all persons wishing to take a higher-school course in physics independently. *Contents.* **Mechanical and Thermal Motion.** The Fundamental Law of Mechanics. Mechanical Energy. Momentum. Rotation of a Rigid Body. Vibrations. Travelling Waves. Standing Waves. Acoustics. Temperature and Heat. Thermodynamic Processes. Entropy. Kinetic Theory of Gases. Processes of Transition to Equilibrium. **Electromagnetic Fields.** Electric Fields. Magnetic Fields. Electromagnetic Fields. Maxwell's Equations. Energy Transformations in Electromagnetic Fields. Electromagnetic Radiation. Propagation of Electromagnetic Waves. Interference Phenomena. Scattering. Diffraction of X-Rays by Crystals. Double Refraction. The Theory of Relativity. The Quantum Nature of a Field. **Structure and Properties of Matter.** Streams of Charged Particles. The Wave Properties of Microparticles. Atomic Structure. Molecules. Atomic Nuclei. Nuclear Transformations. Atomic Structure of Bodies. Phase Transformations. Deformations of Bodies. Dielectrics. Magnetic Substances. Effect of Electron Structure on Properties of Bodies.

Problems in Theoretical Physics

L. G. GRECHKO, V. I. SUGAKOV, O. F. TOMASEVICH,
A. M. FEDORCHENKO

A study aid for university students, based on L. D. Landau and E. M. Lifshitz's *Course of Theoretical Physics*. Compiled by members of the Physics Faculty of Kiev University. Half the volume of the book consists of answers, solutions, and pointers for the 575 problems set.

Contents. Classical Mechanics (122 problems). Electrodynamics (147 problems). Quantum Mechanics (135 problems). Statistical Physics and Thermodynamics (171 problems).

CONTENTS

PREFACE

This textbook is based on lectures which the author has been reading to the students of the Semiconductor Materials and Devices Faculty of the Moscow Institute of Steel and Alloys. The course in Physics of Semiconductors is read after the students become familiar with such courses as "Crystallography", "Quantum Mechanics", etc. This is why this textbook does not treat any questions pertaining to crystal lattice structure, types of atomic bonds in the lattice, etc. Students' familiarity with the principal concepts of quantum mechanics enables its methods to be used to obtain a more rigorous solution of problems arising in the course of formulating various aspects of semiconductor physics. This applies, practically, to every section of the course, and first of all to the fundamentals of the energy band structure and to the charge carrier transition processes involving the interaction with lattice defects, phonons and photons.

Experience has shown that the rather high standards employed in formulating the material do not present undue difficulties to the students. The understanding is substantially facilitated by the presence of practically all of the intermediate calculations. The results of the calculations are sometimes illustrated by experimental data.

The use of the methods of the theory of groups simplifies the handling of many of the problems involved. However, the author bearing in mind that the theory of groups is as a rule not studied in technical colleges felt he had no right to use the theory of groups approach in formulating the material contained in the book limiting himself to an Appendix containing the fundamentals of this theory.

Treating Semiconductor Physics as a separate discipline the author shied away from anything that dealt with the operation of concrete semiconductor devices.

The author uses this occasion to thank everyone who co-operated in the work that led to the publication of this edition of Semiconductor Physics.

The author.

Chapter I

INTRODUCTION. ELECTRON THEORY OF CONDUCTIVITY

1. ELECTRON THEORY OF CONDUCTIVITY. OHM'S LAW

Many concepts of modern physics are based on the electron theory of metals. This is true especially of such concepts as electric conductivity and carrier mobility.

The electron theory of metals developed in the XIX century presumes the electron gas to be in thermal equilibrium with the crystal lattice. The electron gas is presumed to be analogous to the ideal gas of molecular physics in that *it occupies no volume, and there is no interaction between the electrons.* The state of motion of each particle is described by six quantities: the three co-ordinates x, y, z and three velocity components v_x, v_y, v_z (or momentum components p_x, p_y, p_z), or by two vectors **r** and **v** (or **p**). The assumption of a negligible volume is apparently a correct one, since, according to the classical theory, the radius of an electron $r_0 \approx 10^{-15}$ m and the volume $V_0 \approx 10^{-45}$ m^3. Taking the number of electrons per unit volume to be $n \approx 10^{28}$ m^{-3}, we calculate the relative volume of the electrons proper b to be $b = n V_0 \approx 10^{-17}$ of the volume of the body. However, the assumption of no interaction between the electrons seems to be absolutely unjustified. In fact, the charge of the electron is $e = 1.6 \cdot 10^{-19}$ C, and the force of interaction between two electrons at a distance of 10^{-10} m is $2 \cdot 10^{-8}$ N. The acceleration of an electron that would result from the application of this force would amount to $2 \cdot 10^{22}$ m/s^2. The energy of Coulomb interaction between two electrons at a distance of $r = 10^{-10}$ m is about 14 eV.

The total energy of Coulomb interaction (repulsion) of all the electrons should reach enormous positive values. The experiment, on the other hand, proves the energy of the electrons in metals to be negative (in relation to the energy of an electron at an infinite distance from the metal). This is due to the fact that besides the forces of repulsion between the electrons there are the forces of Coulomb attraction acting between the electrons and the atomic nuclei. The force and energy of this interaction are of the same order of magnitude as those of the electron-electron interaction. Moving in the composite field of all electrons and nuclei each electron experiences both attraction and repulsion. Those two types of interaction together bring about the "apparent independence" of motion of the individual electron. As will be shown in Chapter II, the laws of quantum mechanics do, in fact, allow for the electrons to be considered as noninteracting particles.

The electrons move in the crystal at random. In the course of motion they "collide" with the lattice ions, and this changes their velocities both in *modulus* and *direction*. The change in velocity modulus of the electron is connected with the change in its kinetic energy. In conditions of thermal equilibrium the temperature of the electron gas should be equal to the temperature of the lattice ions. This means that, on the average, there is no energy transfer from the electrons to the lattice or vice versa.

But should the temperature of the electron gas be changed, the temperature of the lattice, too, would change as a result of the exchange of energy between the electrons and the ions. This fact is important for explaining the conductivity of metals and semiconductors and will be made use of below.

Because of the random nature of the scattering of electrons after their collisions with the lattice, the velocity and displacement of a given electron averaged over a long period of time considered as vectors should be equal to zero. The conditions are the same for all the electrons; therefore this is true of *every* electron. Since the mean displacement of the electrons taking part in random (thermal) motion is zero such random motion cannot result in an electric current which describes the transport of a charge across some cross section. To establish a current a directional motion of the electrons is needed. This may be initiated by various factors: electric field, temperature gradient, nonuniform illumination, etc.

If an electric field of intensity **E** is established in a metal the electrons will be accelerated by this field. The acceleration of the electron in this field is

$$\mathbf{a}=\frac{e}{m}\mathbf{E}. \qquad (1.1)$$

In time t the electron attains the velocity

$$\mathbf{v}=\mathbf{a}t=\frac{et}{m}\mathbf{E}, \qquad (1.2)$$

directed against the field. If the initial velocity of the electron is $\mathbf{v}_T$, its speed at the moment t will be equal to

$$\mathbf{a}t+\mathbf{v}_T=\frac{et}{m}\mathbf{E}+\mathbf{v}_T \qquad (1.3)$$

As we see from here, the electron velocity component in the direction of the field decreases, and that against the field increases. As a result the electron ensemble attains an appropriate directional velocity. The electrons moving at random at the same time take part in the motion against the field. *The directional motion of the electron ensemble in an electric field is termed drift, and the velocity of directional motion is termed drift velocity and is designated by* $\mathbf{v}_d$. Acted upon by the field **E** the electron in time t will be displaced to a distance

$$\mathbf{l}(t)=\frac{et^2}{2m}\mathbf{E}. \tag{1.4}$$

The classical electron theory presumes the change in velocity to result from an instant interaction of the electron with the lattice (with the atoms or ions of the lattice). In other words, it is assumed that the electron-lattice interaction is analogous to the impact phenomenon in mechanics. Between two collisions the electron moves as a particle free from the action of the field of the lattice and of the other electrons.

To describe the motion of the electrons the concepts of *mean free transit time* (the mean time between two collisions) τ and of *mean free path* l are introduced.

The mean free path l is related to the mean free transit time τ by the expression

$$l=v_T\tau. \tag{1.5}$$

Here v_T is the mean velocity of thermal motion of the electrons, i.e. the mean value of the velocity modulus.

Let us determine the mean electron drift velocity in an electric field. If at $t=0$ the velocity of directional motion of the electron is zero, at $t=\tau$ it will be equal to

$$\mathbf{a}\tau=\frac{e\tau}{m}\mathbf{E}. \tag{1.6}$$

The drift velocity will be equal to the mean velocity of directional motion, i.e. to the half-sum of the initial and the final velocities

$$\mathbf{v}_d=\frac{0+\mathbf{a}\tau}{2}=\frac{e\tau}{2m}\mathbf{E}. \tag{1.7}$$

It follows from (1.7) that the mean velocity of directional motion is proportional to the electric field intensity **E**. *The coefficient in the relation between the drift velocity and field intensity is termed electron mobility and is denoted by the letter* μ:

$$\mu=\frac{e\tau}{2m};\quad \mathbf{v}_d=\mu\mathbf{E}. \tag{1.8}$$

Numerically, electron mobility is equal to electron drift velocity in an electric field of unit intensity.

If electron concentration is n, then per unit time a charge will pass through a unit cross section which is contained in a parallelepiped of unit base and height equal to $1 \cdot v_d$. The charge passing through unit cross section per unit time is termed current density $\mathbf{j}$, and we may write

$$\mathbf{j} = en\mathbf{v}_d = en\,\mu\mathbf{E} = \sigma\mathbf{E}. \tag{1.9}$$

Equation (1.9) is the expression of Ohm's law in differential form. With the aid of (1.9) and (1.8) we obtain the electric conductivity

$$\sigma = en\mu \tag{1.10}$$

and

$$\sigma = \frac{e^2 n\tau}{2m}. \tag{1.11}$$

Expression (1.11) was first obtained by Drude. Expressing τ out of (1.5) we may write (1.11) in the form

$$\sigma = \frac{e^2 nl}{2mv_T}. \tag{1.12}$$

Ohm's law is valid as long as the electric field does not change electron concentration n or mobility μ. However, as the field $\mathbf{E}$ increases, the concentration and mobility of the electrons may change under its influence. Here, for example, is what happens to mobility.

In deducing Ohm's law we made the assumption that the energy of directional motion of the electron as a result of every collision is fully transferred to the lattice. In case of weak electric fields the drift velocity is much smaller than the thermal velocity and for this reason τ is independent of the field intensity $\mathbf{E}$. But as the field increases the drift velocity grows to become comparable to the thermal velocity, and this will lead to a decrease in the free transit time, since now

$$\tau = \frac{l}{v_T + v_d} \tag{1.13}$$

The electron mobility and the conductivity of the metal will decrease accordingly. The critical field E_{cr} at which the effect sets in will be the smaller, the smaller is v_T, i.e. the temperature of the body, and the greater is the electron mobility in weak fields $\mathbf{E}$.

There is another way in which the concept of the free transit time may be interpreted. If at some moment we turn the electric field off, the electron ensemble will continue its directional motion until it, as a result of collisions, transfers all its kinetic energy accumulated in the field to the lattice.

This directional motion will cease after a mean time τ (for all the electrons). After that the electrons will return to the state of random thermal motion.

We see from here that *collisions tend to bring the electron ensemble into thermal equilibrium with the lattice, while the electric field tends to disturb that equilibrium. The passage of a system from the state of non-equilibrium to the state of equilibrium is termed relaxation process or relaxation, and the time in which equilibrium previously disturbed is recovered is termed relaxation time.* Thus, we may say that the *free transit time is actually the relaxation time.*

The SI unit for electrical conductance is siemens (S), and that for specific conductance or conductivity, S/m. The dimensions of mobility in SI may be obtained from (1.8)

$$[\mu]=(\mathrm{A\cdot s^2})\ \mathrm{kg^{-1}}.$$

Mobility may also be expressed as velocity in a unit field, i.e. as a quantity having the dimensions m/s: V/m $=\mathrm{m^2/(V\cdot s)}$. In practice the most widely used units are those that do not belong to the system, i.e. cm, V, s, and mobility is measured in $\mathrm{cm^2/(V\cdot s)}$ and conductivity in $\mathrm{ohm^{-1}\cdot cm^{-1}}$ ($\Omega^{-1}\cdot\mathrm{cm^{-1}}$). Evidently,

$$1\ \mathrm{m^2/(V\cdot s)}=10^4\ \mathrm{cm^2/(V\cdot s)},$$

$$1\ \mathrm{S/m}=10^{-2}\frac{1}{\mathrm{ohm\cdot cm}}\left(\frac{1}{\Omega\cdot\mathrm{cm}}\right).$$

Summary of Sec. 1

1. Main features of the classical electron theory of conductivity are as follows:

(a) Electrons make up an ideal (electron) gas and take part in random thermal motion which is described by the mean free path l and the mean free transit time τ.

(b) Electrons exchange energy and momentum with lattice ions and this causes thermal equilibrium between the electron gas and the lattice to be maintained.

(c) The electric field imparts directional velocity to the electrons thereby causing electric current to flow.

2. The current density is proportional to electric field intensity

$$\mathbf{j}=\sigma\mathbf{E}. \qquad (1.1s)$$

3. Conductivity is related to electron concentration n and mobility μ

$$\sigma=en\mu. \qquad (1.2s)$$

4. Mobility is determined by the behaviour of the electron and by the nature of its random motion in weak fields

$$\mu=\frac{e\tau}{2m}=\frac{el}{2mv_T}. \qquad (1.3s)$$

5. Numerically, the mobility is equal to the electron drift velocity in a unit field

$$\mu = \frac{v_d}{E}. \tag{1.4s}$$

6. Ohm's law ceases to be valid in strong electric fields.
7. Free transit time may be regarded as relaxation time.

2. MEAN-FREE TIME AND FREE-PATH DISTRIBUTION FUNCTIONS

The above expressions for mobility and conductivity which were obtained in the assumption that mean transit times for all the electrons are equal should be rewritten to take account of the fact that the values of mean free time are not the same for all the electrons but vary from 0 to ∞. To do this one should know the probabilities of definite mean free time values assumed to be random quantities. Let us find the distribution function for mean free times.

Let us make the assumptions that

(1) *the probability for an electron to be involved in a collision (scattering) during the time interval dt is proportional to its duration;*

(2) *collision probability per unit time is independent of time.*

These two assumptions suffice for the mean-free time distribution function to be obtained. Let us denote the probability of a particle to move without collision during the time from t to $t+dt$ by

$$dw = dw\,(dt). \tag{2.1}$$

The quantity $w(t)$ is the probability of free motion during the time interval $(t, t+1)$, and $w(t+dt)$, the probability of free motion during the time from $t+dt$ to $t+dt+1$. The quantity $w(t+dt)$ can be expressed in two different ways. On the one hand,

$$w(t+dt) = w(t) + \frac{dw}{dt}dt. \tag{2.2}$$

On the other hand, the fact of free motion of the electron during the time $t+dt$, the event C, may be regarded as a product of two events—one A which is the fact of free motion during time t, and the other B, the fact of free motion during dt, so that

$$C = AB. \tag{2.3}$$

The probability of the product of two events is equal to the product of the probability of the first by the conditional probability of the second:

$$w(C) = w(A)\,w\left(\frac{B}{A}\right) = w(B)\,w\left(\frac{A}{B}\right) \tag{2.4}$$

However, since the event A is independent of the event B, it follows that $w\left(\frac{A}{B}\right)=w(A)$. Besides, the event B turns out to be independent of A.* Therefore

$$w(t+dt)=w(t)\cdot dw(dt). \tag{2.5}$$

The probability of free motion during the time dt may be expressed through the scattering probability within the same time interval. Let us denote the probability of scattering (collision) per unit time by a. Then the scattering probability during time dt will be equal to adt, and the probability of free motion will be equal to $1-adt$, i.e.

$$dw(dt)=1-adt. \tag{2.6}$$

Taking into account (2.2), (2.5) and (2.6), we may write

$$w(t+dt)=w(t)+\frac{dw}{dt}dt=w(t)[1-a\,dt]\doteq w(t)-w(t)\,a\,dt, \tag{2.7}$$

and thus obtain a differential equation for the function $w(t)$

$$\frac{dw}{dt}=-wa. \tag{2.8}$$

Solving equation (2.8), we obtain

$$w(t)=ce^{-at} \tag{2.9}$$

Integration constant c is determined from the normalization condition

$$\int_0^\infty w(t)\,dt=1=c\int_0^\infty e^{-at}dt=\frac{c}{a}, \tag{2.10}$$

whence

$$c=a. \tag{2.11}$$

Accordingly, *the normalized free transit time distribution function assumes the form*

$$w(t)=ae^{-at}. \tag{2.12}$$

* This statement asserting the independence of the event B from the event A usually provokes decisive protest from the students who readily accept the independence of A from B as an obvious fact. The usual argument is that if the particles are scattered during the time before t, there will be no motion in the time from t to dt. But in this case not only the event B will be non-existent, but C as well, and even A. At the same time the theorem is formulated for the event $AB=C$, the product of two events being tantamount to the realization of both events which are the multiplicands in the product.

Let us find the mean free transit time $\langle t\rangle$:

$$\langle t\rangle=\int_0^\infty tw(t)\,dt=\int_0^\infty tae^{-at}\,dt=\frac{1}{a}, \tag{2.13}$$

If we denote the mean free transit time by τ, $\langle t\rangle=\tau$, we will obtain from (2.13) that *the collision probability per unit time is proportional to the mean free transit time:*

$$a=\frac{1}{\tau}=\tau^{-1}. \tag{2.14}$$

The distribution function normalized to unity may be written in the form

$$w(t)=\frac{1}{\tau}e^{-\frac{t}{\tau}}. \tag{2.15}$$

In the same way *the distribution function for the free path x,*

$$w(x)=\frac{1}{l}e^{-\frac{x}{l}}, \tag{2.16}$$

may be obtained. Here l is *the mean free path.*

Distribution functions obtained above are valid for quite general cases. Let us apply them to the case of electrons moving in an electric field. We will be concerned only with the directional motion of the electrons against the field. Suppose the x-axis coincides with the direction of the field. The electron moving freely in the field during the time t attains the velocity

$$\mathbf{v}(t)=\frac{et}{m}\mathbf{E} \tag{2.17}$$

and travels the distance x:

$$x=\frac{eE}{2m}t^2,\quad \mathbf{x}=\frac{e\mathbf{E}}{2m}t^2. \tag{2.18}$$

The mean drift velocity is

$$\mathbf{v}_d=\int_0^\infty \mathbf{v}(t)\,w(t)\,dt=\frac{e\mathbf{E}}{m}\int_0^\infty te^{-\frac{t}{\tau}}\frac{dt}{\tau}=\frac{e\tau}{m}\mathbf{E}, \tag{2.19}$$

and the mean distance travelled against the field

$$\mathbf{l}=\int_0^\infty \mathbf{x}(t)\,w(t)\,dt=\frac{e\mathbf{E}}{2m}\int_0^\infty t^2e^{-\frac{t}{\tau}}\frac{dt}{\tau}=\frac{e\tau^2}{m}\mathbf{E}. \tag{2.20}$$

It follows that *the drift velocity* is equal to

$$\mathbf{v}_d = \frac{e\tau}{m}\mathbf{E} = \mu\mathbf{E}, \tag{2.21}$$

where

$$\mu = \frac{e\tau}{m} \tag{2.22}$$

is the *mobility*.

Expression (2.22) for mobility differs from (1.8) obtained above by the factor 2 which takes into account greater contribution to the collective motion of the electrons with greater free transit times. This is especially evident from the expression for the mean displacement:

$$\langle \mathbf{x} \rangle = \frac{e\tau^2}{m}\mathbf{E}\left(\text{but not } \frac{e\tau^2}{2m}\mathbf{E}\right), \tag{2.23}$$

since always

$$\langle t^2 \rangle \geqslant \langle t \rangle^2; \tag{2.24}$$

and in this case

$$\langle t^2 \rangle = 2\langle t \rangle^2 = 2\tau^2. \tag{2.25}$$

To conclude the section a simple but quite essential note should be made: τ is the mean free time, i.e. the mean time between two collisions. It is determined by the mean free path and the total velocity of the particle but not by its drift velocity. The total velocity depends on the kinetic energy of the particle. It follows that *mean free time is a function of the particle energy*. Therefore, to calculate the drift velocity, as well as various other quantities, the averaging over the free transit times should be done with the electron energy distribution function in mind, as it will be done in Chapter IV.

Summary of Sec. 2

1. The following assumptions suffice to obtain the function $w(t)$: (a) the scattering probability during the time dt is proportional to dt; (b) there exists a time-independent scattering probability per unit time equal to a.

2. Because of random nature of collisions, free transit times may assume various values. The probability of free transit time lying within the interval t, $t+1$ is equal to

$$w(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}}. \tag{2.1s}$$

3. Mean free time τ and the probability of collisions per unit time are related by the expression

$$a\tau = 1. \tag{2.2s}$$

4. The probability of the mean free path to lie within the bounds $x, x+1$ is equal to

$$w(x) = \frac{1}{l} e^{-\frac{x}{l}}. \tag{2.3s}$$

5. The relation of the mean free path l and the probability of scattering per unit path b is given by the expression

$$bl = 1.$$

6. If one takes into account the statistical spread of free transit times, as well as the fact that because of a quadratic dependence of the path on transit time, greater free transit times contribute more to conductivity than the smaller, one obtains

$$\mu = \frac{e\tau}{m}. \tag{2.4s}$$

3. ELECTRON DISTRIBUTION FUNCTION. MEAN VALUES OF PHYSICAL QUANTITIES

The space of co-ordinates x, y, z and momenta p_x, p_y, p_z is termed the phase space. Consider an element of volume in the phase space

$$d\Gamma = dx\,dy\,dz\,dp_x\,dp_y\,dp_z = d\tau_r\,d\tau_p \tag{3.1}$$

containing the point with the co-ordinates $(x, y, z; p_x, p_y, p_z)$. The number dN of electrons in this element of volume $d\Gamma$ should be proportional to $d\Gamma$ provided $d\Gamma$ is sufficiently small. Besides, it should depend on the point $(\mathbf{r}, \mathbf{p})$ of the phase space, this dependence being expressed by some function $F(\mathbf{r}, \mathbf{p}, t)$, so that

$$dN = F(\mathbf{r}, \mathbf{p}, t)\, d\Gamma. \tag{3.2}$$

Integrating over the entire phase space, we obtain the total number of electrons

$$N = \int_{(V_\Gamma)} F(\mathbf{r}, \mathbf{p}, t)\, d\Gamma. \tag{3.3}$$

The function $F(\mathbf{r}, \mathbf{p}, t)$ determines the number of electrons per unit volume of the phase space. An N times smaller function $f(\mathbf{r}, \mathbf{p}, t)$ may be introduced instead:

$$f(\mathbf{r}, \mathbf{p}, t) = \frac{1}{N} F(\mathbf{r}, \mathbf{p}, t). \tag{3.4}$$

The function $f(\mathbf{r}, \mathbf{p}, t)$ *determines the probability of one electron occupying a unit volume of the phase space.* The equation

$$\int\limits_{(V_\Gamma)} f(\mathbf{r}, \mathbf{p}, t)\, d\Gamma = 1 \tag{3.5}$$

is the normalization condition.

The integration is done over the entire volume V of the crystal and over all possible values of the momentum. The normalization condition is independent of time; the distribution function, on the othez hand, is generally time-dependent. With the help of the distribution function the mean values of an arbitrary co-ordinate- and momentum-dependent physical quantity can be calculated. Suppose α is a quantity which may depend on the state of electron motion: $\alpha = \alpha(\mathbf{r}, \mathbf{p}, t)$. Let us take an element $d\Gamma$ of the phase space. It contains

$$dN = Nf(\mathbf{r}, \mathbf{p}, t)\, d\Gamma \tag{3.6}$$

electrons, each of which possesses the quantity $\alpha(\mathbf{r}, \mathbf{p}, t)$.

The total value of the physical quantity for dN electrons will be

$$\alpha\, dN \tag{3.7}$$

If we integrate (3.7) over the entire phase space, we will obtain the sum-total value of the quantity α for all the electrons. If we further divide this value by N, we will obtain the mean value

$$\langle\alpha\rangle = \frac{1}{N}\int\limits_{(V_\Gamma)} \alpha\, dN = \int\limits_{(V_\Gamma)} \alpha(\mathbf{r}, \mathbf{p}, t) f(\mathbf{r}, \mathbf{p}, t)\, d\Gamma. \tag{3.8}$$

If (3.7) is integrated only over the momentum and divided by $N\, d\tau_r$, the mean value of $\overline{\alpha(\mathbf{r})}^p$ for the electrons in the vicinity of the point $\mathbf{r}$ will be obtained.

One specific case important for practical purposes should be noted. Suppose the distribution function is independent of the co-ordinates. In this case integration of relation (3.5) over the co-ordinates yields the volume of the body V. For this reason integration of the distribution function over the momentum yields the reciprocal volume. If α, too, is independent of the co-ordinates $\mathbf{r}$, averaging over the momentum alone will yield $\overline{\alpha}^p = \frac{\langle\alpha\rangle}{V}$, where V is the volume of the body.

Current density may be expressed in terms of the distribution function in the following way:

$$\mathbf{j} = en\langle\mathbf{v}\rangle = enV\int\limits_{-\infty}^{\infty} \mathbf{v} f(\mathbf{r}, \mathbf{p}, t)\, d\tau_p = en\mathbf{v}_d. \tag{3.9}$$

If $f(\mathbf{r}, \mathbf{p}, t) = f(\mathbf{r}, -\mathbf{p}, t)$, which means that the probabilities of the electron motion with velocities $\mathbf{v}$ and $-\mathbf{v}$ are equal, the mean velocity $\langle\mathbf{v}\rangle$, by reason of $\frac{\mathbf{p}}{m} f(\mathbf{r}, \mathbf{p}, t)$ being an odd function, turns zero when the limits of integration are symmetrical.

The motion of electrons in a metal in a state of *thermodynamical equilibrium* is described by an even (symmetrical) distribution function. This leads to a well-established fact that *thermal motion is incapable of generating current*. To generate current the conditions should be established in which the symmetry of the distribution function is disturbed and, consequently, some velocity directions become more probable than the others. This results in the electron ensemble as a whole moving in space. Directional motion of the electrons is superimposed on random motion. There may be various causes for this directional motion: electric field, temperature gradient, inhomogeneous illumination, etc.

Generally, electron distribution function should be obtained from the Boltzmann kinetic equation which will be discussed together with kinetic phenomena in semiconductors.

For a system in a state of equilibrium the following expression is derived in quantum statistics ($f = f_0$):

$$f_0(E, T) = \frac{1}{e^{\frac{E-F}{kT}} + 1}. \tag{3.10}$$

In this expression E is the total energy characteristic of the electron state, T, the absolute temperature, k, Boltzmann's constant, F is termed the Fermi energy (level).

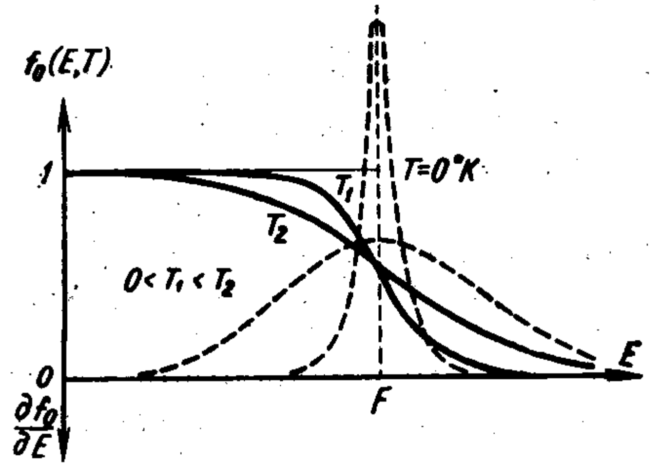


Fig. 1. The Fermi-Dirac function and its energy derivative at various temperatures

The quantity $f_0(E, T)$ is the probability of the state with the energy E being occupied by an electron. The expression $f_0(E, T)$ is termed the Fermi-Dirac function. We see from here that the probability of an electron occupying the state with the energy E depends on the energy E and on temperature T. Besides, this probability depends also on the Fermi energy the physical meaning of which will be evident from the consideration of some of the simplest cases.

Suppose $T \to 0$. Then

$$\lim_{T\to 0} f_0 = \begin{cases} 1 \text{ for } E < F, \\ 0 \text{ for } E > F. \end{cases} \qquad (3.11)$$

When $E = F$ the function is not defined. Moreover, it becomes discontinuous. The graph of $f_0(E, T)$ is shown in Fig. 1. We see from here that when $T = 0$ all states for which $E < F$ are occupied while all states for which $E > F$ are absolutely empty, not occupied by the electrons. As will be shown below, the state $E = F$ may be assumed to be occupied with a probability of 0.5.

In this case *the physical meaning of the Fermi energy* is quite obvious—it *is the maximum energy of electrons in a metal at absolute zero temperature.* In addition it may be said that *the Fermi level separates the occupied energy states from the empty ones.*

Generally, *the Fermi energy is the Gibbs thermodynamic potential per particle. The other name for F is the chemical potential.* The Fermi energy is numerically equal to work which must be expended to add one particle to the system.

Let us now consider the case $T \neq 0$. First and foremost we discover that for any $T \neq 0$ the probability of the state $E = F$ being occupied remains equal to 0.5. This enables us to define the value of $f_0(E = F, 0) = 0.5$. For $E \ll F$ $f_0(E, T) \approx 1$. For $E \gg F$ the Fermi-Dirac function may be written as follows:

$$f_0(E, T) \approx e^{\frac{F}{kT}} e^{-\frac{E}{kT}}. \qquad (3.12)$$

The probability $f_0(E, T)$ experiences a sharp change in the several kT wide interval centred around $E = F$. Suppose $E = F + \xi\, kT$, where ξ is a variable energy expressed in units of kT and measured from the Fermi level

$$\xi = \frac{E - F}{kT}. \qquad (3.13)$$

In this case

$$f_0 = \frac{1}{e^{\xi} + 1}. \qquad (3.14)$$

Table 1 shows the value of $f_0(\xi)$ for several values of ξ.

The graph of $f_0(\xi)$ in Fig. 2 shows the Fermi-Dirac function changing sharply from 0.88 to 0.12 when ξ changes from -2 to 2.

Table 1

ξ	-5	-4	-3	-2	-1	0	1	2	3	4	5
f_0	0.994	0.980	0.954	0.882	0.730	0.500	0.270	0.118	0.048	0.018	0.007

In other words, the *distribution function $f_0(E, T)$ experiences essential changes when the energy is changed from $F-2kT$ to $F+2kT$.* For $\xi<-2$, $f_0\approx 1$; for $\xi>2$, f_0 behaves like an exponential function. The energy interval within which $f_0(E, T)$ changes sharply depends on temperature: when $T\rightarrow 0$ it contracts to zero. To assess the changes in the distribution function it is convenient to make use of its derivative with respect to energy:

$$f'_{0E}=\frac{\partial f_0}{\partial E}=\frac{\partial f_0}{\partial \xi}\frac{\partial \xi}{\partial E}=-\frac{1}{2kT}\frac{1}{1+\cosh\xi}, \qquad (3.15)$$

where $\cosh\xi=\frac{e^{\xi}+c^{-\xi}}{2}$.

When $T\rightarrow 0$ $\cosh\xi\rightarrow\infty$ everywhere except $E=F$ ($\xi=0$). Therefore when $T\neq 0$ f'_{0E} is not equal to zero only in the near vicinity

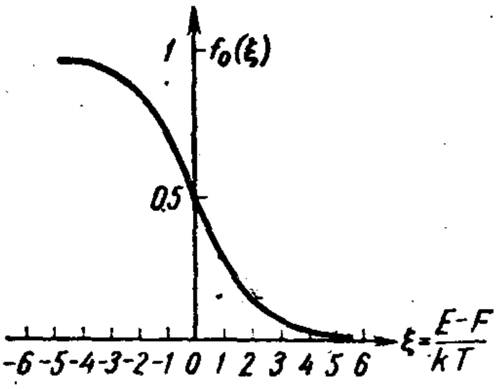


Fig. 2. The Fermi-Dirac function. The energy is expressed in kT and measured from the Fermi level $\left(\xi=\frac{E-F}{kT}\right)$

of the point $\xi=0$. When $T\rightarrow 0$ this vicinity contracts to a single point $\xi=0$ so that

$$-\frac{\partial f_0(E, 0)}{\partial E}=\begin{cases}\infty & \text{for } E=F,\\ 0 & \text{for } E\neq F.\end{cases} \qquad (3.16)$$

Simultaneously

$$-\int_{-\infty}^{\infty} f'_{0E}(E, 0)\,dE=1. \qquad (3.17)$$

Expression (3.17) is valid for any temperature $T\geqslant 0$.

We see from here that $-f'_{0E}(E, 0)$ is the Dirac δ-function.

When $T > 0$ $f'_{0E}(E, T)$ will approach the δ-function the closer the lower is the temperature. The dashed line in Fig. 1 represents the graph of $f'_{0E}(E, T)$ for various temperatures. *Quantum systems which are described by the Fermi-Dirac function are termed degenerate.*

For large values of ξ f_0 may be represented by the exponential function

$$f_0(\xi) = \frac{1}{e^{\xi}+1} \approx e^{-\xi}, \tag{3.18}$$

or

$$f_0(E, T) \approx e^{\frac{F}{kT}} \cdot e^{-\frac{E}{kT}}. \tag{3.19}$$

The function $f_0'(E, T) = e^{F/kT} \cdot e^{-E/kT}$ bears the name of the classical distribution, or Boltzmann, function. *Systems described by the Boltzmann function, i.e. obeying the laws of classical statistics, are termed non-degenerate.* The difference between the classical distribution function and the quantum-mechanical one lies not only in the form but carries a deeper meaning which implies, for instance, the possibility of separate determination, in classical physics, of the kinetic and potential energies, while in quantum mechanics, where only the total energy may be determined, this is impossible.

To calculate mean values of physical quantities with the aid of distribution functions $f_0(E, T)$ it is necessary to know how the energy of the particles depends on their velocity (or the momentum $\mathbf{p}$). Suppose that for the gas of free electrons

$$E = \frac{m\mathbf{v}^2}{2} = \frac{\mathbf{p}^2}{2m} = \frac{1}{2m}(p_x^2 + p_y^2 + p_z^2). \tag{3.20}$$

Since the energy is independent of the momentum direction, an element of volume in the momenta space $d\tau_p = dp_x \, dp_y \, dp_z$ is more conveniently expressed in spherical co-ordinates

$$d\tau_p = p^2 dp d\Omega = p^2 \, dp \sin\theta \, d\theta \, d\varphi. \tag{3.21}$$

Integration over the angles θ, φ yields the factor 4π.

Let us calculate the mean value of the electron energy for the case of Boltzmann's statistics:

$$\langle E \rangle = \frac{4\pi V \int\limits_0^\infty \frac{p^2}{2m} e^{-\frac{p^2}{2mkT}} p^2 \, dp}{4\pi V \int\limits_0^\infty e^{-\frac{p^2}{2mkT}} p^2 \, dp} \tag{3.22}$$

The denominator in (3.22) is due to the fact that the Maxwell distribution function in (3.22) has not been normalized. Integration over the co-ordinates of which in this case the distribution function is independent yields the volume V of the body.

It will be more convenient to introduce a new variable x satisfying the relations

$$\frac{p^2}{2mkT}=x;\quad p^2=(2mkT)\,x;\quad dp=\frac{1}{2}(2mkT)^{1/2}\,\frac{dx}{x^{1/2}}. \tag{3.23}$$

In this case

$$\langle E\rangle=\frac{(kT)\int\limits_0^\infty x^{\frac{3}{2}}\,e^{-x}\,dx}{\int\limits_0^\infty x^{\frac{1}{2}}\,e^{-x}\,dx}=\frac{3}{2}\,kT. \tag{3.24}$$

In the calculation we took account of the fact that

$$\int\limits_0^\infty x^{\frac{3}{2}}e^{-x}\,dx=\frac{3}{2}\int\limits_0^\infty x^{\frac{1}{2}}e^{-x}\,dx. \tag{3.25}$$

This is easily proved by means of integration by parts, the last integral being equal to $\frac{\sqrt{\pi}}{2}$ (Poisson's integral). In the process we also found the normalizing coefficient for the Boltzmann function. This function normalized to unity in the process of integration over the phase space has the form

$$f_0(E,\,T)=\frac{1}{V}\left(\frac{1}{2\pi mkT}\right)^{\frac{3}{2}}e^{-\frac{p^2}{2mkT}}. \tag{3.26}$$

Let us now, making the same assumptions, find the mean energy of the electrons of a degenerate system

$$f_0(E,\,T)=\frac{1}{e^{\frac{\frac{p^2}{2m}-F}{kT}}+1} \tag{3.27}$$

and

$$\langle E\rangle=\frac{4\pi V\int\limits_0^\infty \frac{p^2}{2m}\cdot\frac{1}{e^{\left\{\left(\frac{p^2}{2m}-F\right)\Big/kT\right\}}+1}\,p^2\,dp}{4\pi V\int\limits_0^\infty \frac{1}{e^{\left\{\left(\frac{p^2}{2m}-F\right)\Big/kT\right\}}+1}\,p^2\,dp} \tag{3.28}$$

Introducing the variable (3.23) and cancelling out equal factors in the numerator and the denominator, we may write

$$\langle E\rangle = kT\frac{\int\limits_0^\infty \frac{x^{\frac{3}{2}}\,dx}{e^{\left(x-\frac{F}{kT}\right)}+1}}{\int\limits_0^\infty \frac{x^{\frac{1}{2}}\,dx}{e^{\left(x-\frac{F}{kT}\right)}+1}}. \tag{3.29}$$

At high temperatures, when $kT \gg F$ and $e^{F/kT} \approx 1$, unity may be neglected as compared to e^x in the integrand already for small x. In this case quantum statistics transforms into classical statistics and

$$\langle E\rangle = \frac{3}{2}kT. \tag{3.30}$$

At very low temperatures the Fermi-Dirac function assumes the form of a rectangular step of unit height, and the integral may be calculated from 0 to $x=\frac{F}{kT}$. Therefore

$$\langle E\rangle = \frac{kT\int\limits_0^{\frac{F}{kT}} x^{\frac{3}{2}}dx}{\int\limits_0^{\frac{F}{kT}} x^{\frac{1}{2}}dx} = \frac{3}{5}F. \tag{3.31}$$

As can be seen from (3.31), the *energy of degenerate electron gas is independent of temperature*. Therefore it does not contribute to the specific heat of the body, and this fully explains the Dulong-Petit law for solids.

If the calculation is carried out more rigorously, the following expression for $\langle E\rangle$ may be obtained:

$$\langle E\rangle = \frac{3}{5}F_0\left[1+\frac{5\pi^2}{12}\left(\frac{kT}{F_0}\right)^2\right]. \tag{3.32}$$

Quantum theory achieved a great success in explaining the independence of specific heat in solids, at normal temperatures, from electron motion, since *according to classical theory the electrons should have contributed* $\frac{3}{2}$ *kT to the specific heat of the metal.*

However, at low temperatures the distribution of the electrons over energy states is determined not by thermal motion but by the wave properties of the electrons and by the Pauli exclusion principle that follows from them. Consequently, $\langle E\rangle = \frac{3}{5}F$ which amounts to several electron-volts, and as a result thermal motion exercises little influence on the energy of the electron gas. At elevated temperatures, when $kT > F_0$, the energy of the electron gas is, however, determined by its temperature.

The condition for the transition from quantum to classical statistics may be set down as follows

$$kT_{deg} = F_0 \tag{3.33}$$

or

$$T_{deg} = \frac{F_0}{k}. \tag{3.34}$$

The temperature T_{deg} is termed degeneracy temperature, since at $T > T_{deg}$ the electron gas is a classical gas and at $T < T_{deg}$ a quantum (degenerate) gas.

Both the Fermi-Dirac (for degenerate semiconductors) and the Boltzmann (for non-degenerate semiconductors) functions will be used to describe the properties of semiconductors.

Subsequently we will have to calculate integrals of the form

$$I = \int_0^\infty \varphi(E) f_0(E, T)\, dE. \tag{3.35}$$

Denoting the antiderivative of the function $\varphi(E)$ by $\psi(E)$ and integrating by parts, we obtain

$$I = \psi(E) f_0(E, T)\Big|_0^\infty - \int_0^\infty \psi(E) \frac{\partial f_0(E, T)}{\partial E}\, dE. \tag{3.36}$$

The first term of (3.36) usually turns zero. For this to be the case it suffices for $\psi(0)$ to be zero and for $\psi(E)$ to increase with the increase of energy more slowly than an exponential function. Assuming these conditions to be fulfilled, we will drop the first term of (3.36). Expand $\psi(E)$ into the Taylor series in the vicinity of the point $E = F$

$$I = -\int_0^\infty \sum_{n=0}^\infty \frac{1}{n!} \cdot \frac{d^n\psi(E)}{dE^n}\Big|_F (E-F)^n \frac{\partial f_0(E, T)}{\partial E}\, dE. \tag{3.37}$$

Going over to a dimensionless variable $\xi = \frac{E-F}{kT}$ and taking account of the relation (3.15), we write

$$I = -\sum_{n=0}^\infty \frac{d^n\psi(E)}{dE^n}\Big|_F (kT)^n \cdot \frac{1}{n!} \int_{-\frac{F}{kT}}^\infty \xi^n \frac{df_0}{d\xi}\, d\xi. \tag{3.38}$$

For the case of a strong degeneracy $-\infty$ may be set instead of the lower limit. As a result, the terms with odd powers will turn zero, and (3.38) will assume the form

$$I=\sum_{r=0}^{\infty}\frac{d^{2r}\psi(E)}{dE^{2r}}\bigg|_F (kT)^{2r}C_{2r}, \tag{3.39}$$

where C_{2r} denote definite integrals of the form

$$C_{2r}=-\frac{1}{(2r)!}\int_{-\infty}^{\infty}\xi^{2r}\frac{df_0}{d\xi}d\xi=\frac{1}{(2r)!}\int_{-\infty}^{\infty}\frac{\xi^{2r}e^{-\xi}}{(1+e^{-\xi})^2}d\xi. \tag{3.40}$$

The coefficients C_{2r} are calculated with the aid of some special functions, for instance, the Riemann zeta function $\zeta(2r)=\sum_{l=1}^{\infty}l^{-2r}$.

$$C_{2r}=2(1-2^{1-2r})\zeta(2r); \tag{3.41}$$

With the aid of $\zeta(2r)$, C_{2r} may be found, i.e.

$$\zeta(2)=\frac{\pi^2}{6};\quad \zeta(4)=\frac{\pi^4}{90};\ \ldots\ ;$$

$$C_0=1;\ C_2=\frac{\pi^2}{6};\ C_4=\frac{7\pi^4}{360};\ C_6=\frac{31\pi^6}{5120};\ \ldots$$

or $$C_0=1;\ C_2\cong 1.64;\ C_4\cong 1.89;\ C_6\cong 1.96,\ \text{etc.}$$

For a limited number of terms of the series we obtain

$$\int_0^{\infty}\varphi(E)f_0(E,T)\,dE=\psi(F)+\frac{\pi^2}{6}(kT)^2\psi''(F)+\frac{7\pi^4}{360}(kT)^4\psi''''(F)+\ldots \tag{3.42}$$

In many cases the number of terms may be limited to two

$$\int_0^{\infty}\varphi(E)f_0(E,T)\,dE\cong\psi(F)+\frac{\pi^2}{6}(kT)^2\varphi'(F). \tag{3.43}$$

The expression (3.43) may, for instance, be used to calculate (3.29). Setting $\psi=\frac{2}{5}E^{5/2}$ and $\psi=\frac{2}{3}E^{3/2}$ in the numerator and denominator of (3.29), we obtain

$$\langle E\rangle=\frac{\frac{2}{5}F^{5/2}+\frac{2}{5}\cdot\frac{5}{2}\cdot\frac{3}{2}\frac{\pi^2}{6}(kT)^2F^{1/2}+\ldots}{\frac{2}{3}F^{3/2}+\frac{2}{3}\cdot\frac{3}{2}\cdot\frac{1}{2}\frac{\pi^2}{6}(kT)^2F^{-1/2}+\ldots}\cong$$

$$\cong\frac{3}{5}F\frac{1+\frac{15}{4}\cdot\frac{\pi^2}{6}\left(\frac{kT}{F}\right)^2}{1+\frac{3}{4}\frac{\pi^2}{6}\left(\frac{kT}{F}\right)^2}\cong\frac{3}{5}F\left[1+\frac{\pi^2}{2}\left(\frac{kT}{F}\right)^2\right]. \qquad (3.44)$$

When $T\rightarrow 0$ (3.44), as may be seen, turns into (3.31). In the same way it may be obtained for constant electron concentration

$$F=F_0\left[1-\frac{\pi^2}{12}\left(\frac{kT}{F_0}\right)^2\right]. \qquad (3.45)$$

Substituting (3.45) into (3.44), we obtain (3.32).

Summary of Sec. 3

1. To describe electron gas a distribution function $f(\mathbf{r}, \mathbf{p}, t)$ which gives the probability of the electron occupying a unit phase volume may be introduced.

2. The mean value $\langle\alpha\rangle$ of any physical quantity α may be found with the aid of (3.8).

3. The quantum-mechanical electron distribution function (3.10) depends only on energy and temperature. It approaches unity for $E<F-2\,kT$ and zero for $E>F+2\,kT$, experiences a sharp change within the energy interval of $\pm 2kT$ centred around the Fermi energy F. Systems which obey the laws of quantum statistics are termed degenerate. As the temperature rises and F becomes smaller than kT the Fermi-Dirac function transforms into the Boltzmann, or classical, distribution function (3.12).

4. The expression for the mean energy of a degenerate electron gas is (3.32), and for a non-degenerate gas—(3.24):

$$\langle E\rangle=\frac{3}{5}F_0\left[1+\frac{5\pi^2}{12}\left(\frac{kT}{F_0}\right)^2\right];\quad \langle E\rangle=\frac{3}{2}kT. \qquad (3.1s)$$

5. The degeneracy temperature T_{deg} of the electron gas is determined by the Fermi energy F_0

$$T_{deg}=\frac{F_0}{kT}. \qquad (3.2s)$$

When $T>T_{deg}$ the gas is classical (non-degenerate), when $T<T_{deg}$—it is quantum (degenerate).

6. For degenerate semiconductors and metals the calculations performed with the Fermi-Dirac function lead to the following re-

sults

$$\int_0^\infty \varphi(E)\, f_0(E, T)\, dE = \psi(E)\, f_0(E, T)\Big|_0^\infty +$$
$$+\psi(F) + \frac{\pi^2}{6}(kT)^2 \varphi'(F) + \frac{7\pi^4}{360}(kT)^4 \varphi'''(F) + \ldots, \qquad (3.3s)$$

where

$$\varphi(E) = \frac{d\psi(E)}{dE}.$$

4. SEMICONDUCTORS. THE CLASSIFICATION OF MATERIALS ACCORDING TO THEIR CONDUCTIVITY

Materials are characterized by different conductivities. Some examples are given in Table 2.

Table 2

Material	σ, S·m⁻¹	Material	σ, S·m⁻¹
Aluminium	3.12×10^7	Diamond	10^{-10}
Gold	4.13×10^7	Ebonite	5×10^{-14}
Copper, drawn	5.62×10^7	Pyrex	1×10^{-12}
Copper, annealed	6.30×10^7	Mica	1.1×10^{-11}
Silver	6.03×10^7	Paraffinized wax	3.3×10^{-17}
Nichrome	9×10^5	Quartz	5×10^{-19}

It follows from Table 2 that the conductivities of such materials as gold, silver, copper are several units times 10^7 S·m^{-1}. The corresponding values for ebonite and amber are of the order of 10^{-14} S·m^{-1}. Materials with conductivities $\sigma \cong (10^7\text{-}10^6)$ S·m^{-1} are usually termed conductors, or metals. Isolators or dielectrics are those with conductivities $\sigma \cong (10^{-8}\text{-}10^{-16})$ S·m^{-1}.

Materials with intermediate conductivities $\sigma = (10^{-8}\text{-}10^6)$ S·m^{-1}, *were termed semiconductors*. However, this definition of semiconductors fails to reveal any conductivity characteristics peculiar to them. Consider, for example, the temperature dependence of conductivity of metals and semiconductors. In case of metals the resistance increases with temperature

$$R(t) = R_0(1+\alpha t) \qquad (4.1)$$

Here R_0 is the resistance at $t=0$°C, $R(t)$, the resistance at t°C, α, the thermal resistance coefficient equal to about 1/273. For metals

$$\alpha = \frac{dR}{dt} = \frac{dR}{dT} > 0.$$

In case of semiconductors the resistance rapidly decreases with the increase of temperature. The empirical formula connecting the resistance and absolute temperature T valid within a limited temperature range is given below:

$$R(T) = R_0 e^{\frac{B}{T}}. \tag{4.2}$$

Here R_0, B are, within that temperature range, constants characteristic of each semiconducting material.

This formula may be rewritten for conductivity

$$\sigma = \sigma_0 e^{-\frac{B}{T}}. \tag{4.3}$$

If we multiply the numerator and the denominator of the exponent by the Boltzmann constant k and denote $kB = E_a$, we will obtain

$$\sigma = \sigma_0 e^{-\frac{E_a}{kT}}. \tag{4.4}$$

The quantity E_a characteristic of a given semiconductor is called its activation energy. Its physical meaning is different for different temperature ranges. Figure 3 shows the dependence of the resistance of a metal and a semiconductor on temperature T. Figure 4 shows the dependence of ln σ on the inverse temperature.

The existence of an activation energy E_a means that to increase the coductivity of a semiconductor one should supply energy to it. Experiment shows that the conductivity of semiconductors increases not only upon heating (i.e. when thermal energy is supplied to the semiconductor) but also upon illumination and upon bombardment by nuclear particles; it changes with applied electric or magnetic fields, upon external pressure changes, etc.

This means that *semiconductors are materials the conductivity of which is strongly dependent on the ambient: temperature, pressure, external fields, illumination, and irradiation by nuclear particles.*

Since at $T \to 0$ in the absence of energy supply the conductivity of non-degenerate semiconductors tends to zero, we may say that *semiconductors are materials that exhibit conductivity only in a state of excitation.* Such a definition makes no principal distinction between semiconductors and dielectrics. At the same time the distinction between semiconductors and metals is quite clear-cut.

It depends on the structure and properties of the semiconductor, how the ambient affects its conductivity. In the same ambient the conductivity of a material in the form of a pure defectless single crystal, or of a single crystal containing defects and impurities, or of a polycrystal, will be different.

To take account of everything said above we may define semiconductors as follows. *Semiconductors are materials the conductivity of which at room temperature ranges between* 10^{-8} *and* 10^{6} $\text{S}\cdot\text{m}^{-1}$ *and depends strongly on the nature and amount of impurities, on the structure of the material and on the ambient: temperature, illumination, electric and magnetic fields, etc.* This definition enables

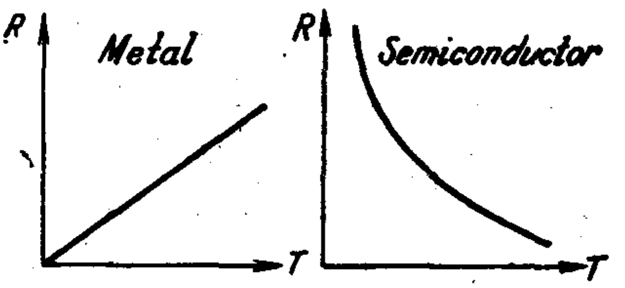


Fig. 3. The temperature dependence of the resistance in metals and semiconductors.

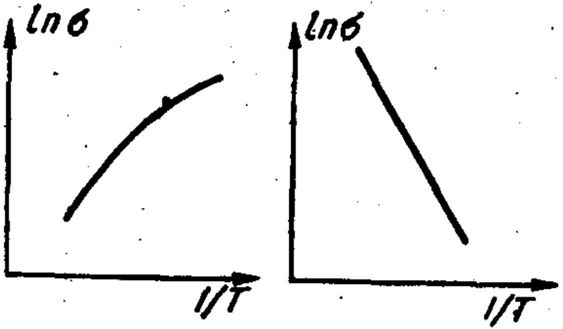


Fig. 4. The dependence of ln σ on the inverse temperature for metals and semiconductors

the distinction between semiconductors and metals to be drawn — the conductivity of metals depends on the ambient much less because in metals the state of conduction is a non-excited* state, while to make semiconductors conductive it is necessary to transform them into an excited (activated) state. The limitation on the possible range of specific conductivities enables one to draw the distinction between simiconductors and dielectrics. Hence, *the distinction between semiconductors and insulators is purely quantitative and, to a large extent, conventional.*

There are two kinds of semiconducting materials: ionic and electronic. In ionic semiconductors the current is carried by the ions of the substance, and for this reason the composition and structure of an ionic semiconductor change as current is passed through it. Such materials are useless for devices designed to transform energy because they will be destroyed by the passing current, and, consequently, they are not treated in this book.

The current in electronic semiconductors is carried by electrons only, there is no transport of matter, and devices made from electronic semiconductors can operate for long periods of time.

There is an enormous number of materials which can be classified as electronic semiconductors. But only some of them are at present of practical importance. However, the number of semiconductors employed in practice grows rapidly with the progress of chemistry and technology of production of pure materials.

12 elements belong to the semiconductor group: boron B, carbon C, silicon Si, phosphorus P, sulphur S, germanium Ge, arsenic As, selenium Se, grey tin Sn, antimony Sb, tellurium Te, iodine I. Figure 5 shows the position of the semiconductors among

Group Period	II	III	IV	V	VI	VII	
II	Be	B	C	N	O		
III		Al	Si	P	S	Cl	
IV		Ca	Ge	As	Se	Br	
V		In	Sn	Sb	Te	I	Xe
VI			Pb	Bi	Po	At	

Fig. 5. The position of atomic semiconductors in the Mendeleyev table

the other elements of Mendeleyev's Periodic Table. Most important atomic semiconductors are at present silicon and germanium.

Many binary compounds of the A^XB^{8-X} type (where A is the element of the group X and B—the element of the group 8—X) exhibit semiconductive properties.

Type A^IB^{VII} compounds include AgCl, CuBr, KBr, LiF, etc. As yet, they are not in great use as semiconductive materials. Type $A^{II}B^{VI}$ compounds include sulphides, tellurides, selenides and oxides of group II metals. The best known among them are the compounds CdS, CdSe, CdTe, ZnS, ZnO, ZnSe, HgTe, HgSe, etc. They will be widely used in the near future.

Among the most important present day semiconductive compounds are the substances of the $A^{III}B^V$ type. They include antimonides, arsenides, phosphides, and nitrides of the sub-group II of group III (aluminium, gallium, indium, and boron).

Type $A^{IV}B^{IV}$ compounds include SiC, SiGe. Besides A^XB^{8-X} type compounds the class of semiconductive substances includes also such compounds as $A^{IV}B^{VI}$ (PbS, PbSe, PbTe), A^IB^{VI} (CuS, CuO, Cu_2O), etc.

Of great interest are the more complex compounds and solid solutions which include substances of the $A^XB_1^{8-X}B_2^{8-X}$, $A_1^XA_2^XB^{8-X}$, $A_1^XA_2^XB_1^{8-X}B_2^{8-X}$ types. GaAsP, InGaSb or ZnCdSeTe may serve as examples of solid solutions.

It will be possible by combining different elements to obtain compounds with properties most suited for specific practical purposes. There is also a vast number of various complex compounds belonging to the semiconductor class.

Besides inorganic substances semiconductors include some organic substances, as well, such as anthracene methylene blue, phthalocyanine, coronene, etc.

The availability of semiconductors with a wide diversity of properties facilitated their widespread use for the manufacture of various devices..

Semiconductor diodes can rectify currents ranging from milliamperes to thousands of amperes, from low to ultra-high frequencies, voltages from a fraction of a volt to kilovolts. Transistors are employed to amplify and generate oscillations in a wide frequency band. Registration of luminous and particle radiation, conversion of radiative and thermal energies into electric energy are all being done with the aid of highly efficient semiconductor receivers and convertors.

Techniques for measuring pressures, magnetic fields, temperature, and radiative energies make wide use of various "gauges" which convert various influences into electric signals.

Semiconductors are being widely used for efficient generation of coherent radiation, for the conversion of electric energy into light. It is difficult even to enumerate all the fields where semiconductors are already being employed.

Every device is based upon some physical processes and phenomena without the knowledge of which it is impossible to use and design new devices correctly.

5. SEMICONDUCTOR CONDUCTIVITY MODELS. THE CONCEPT OF A HOLE

A crystal lattice is a result of atomic interaction. The nature of this interaction is determined by the atoms constituting the crystal. Main part in this interaction is played by the so-called exchange effect in the course of which two atoms exchange electrons and thereby establish forces of attraction between them. This bond is called homopolar since in most cases it exists between similar atoms. Since it is due primarily to valence electrons, it is also called the covalent bond. The strongest bond is established in the case of two electrons with opposite spins. It is evident from here that the covalent bond must display saturation—the addition of a third electron cannot increase the bonding energy since in this case two of the electrons will always have identical spins. This property of the covalent bond follows from the Pauli principle which, as applied to this case, states that two electrons with identical spin projections cannot at the same time occupy the same interatomic space. Diamond, silicon and germanium are examples of substances with covalent bonds.

When two different atoms interact the position of maximum density of the electron cloud (the probability of electron location) can shift closer to one of them, i.e. to the one with the greater number of electrons in its valence shell. In the extreme case the electron cloud will have maximum density around one of the interacting atoms which thereby turns into a negative ion, while the other atom becomes a positive ion. The bond in this case is termed ionic since it may be considered to be the result of Coulomb's attraction of oppositely charged ions. Ionic bond is most clearly displayed in alkalinehalogenide crystals (NaCl, LiF, etc.). Ionic bond is the limiting case for the covalent bond. The chemical bond in such substances as oxides and sulphides of the group II elements is said to be so-and-so per cent ionic. The other limiting case for exchange interaction is the exchange of electrons between any (not just neighbouring) pairs of atoms in the crystal. The electrons belong to the crystal lattice as a whole, they are collectivized. This bond is called metallic for it is in metals that it is most pronounced. It must, however, be remembered that no strict line may be drawn between the three types of bond. The bond in organic crystals is the comparatively weak molecular bond established as a result of the so-called van der Waals forces which are based upon the interaction of induced molecular dipole moments.

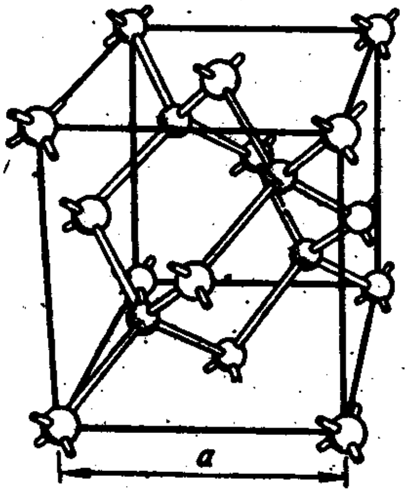


Fig. 6. The diamond lattice

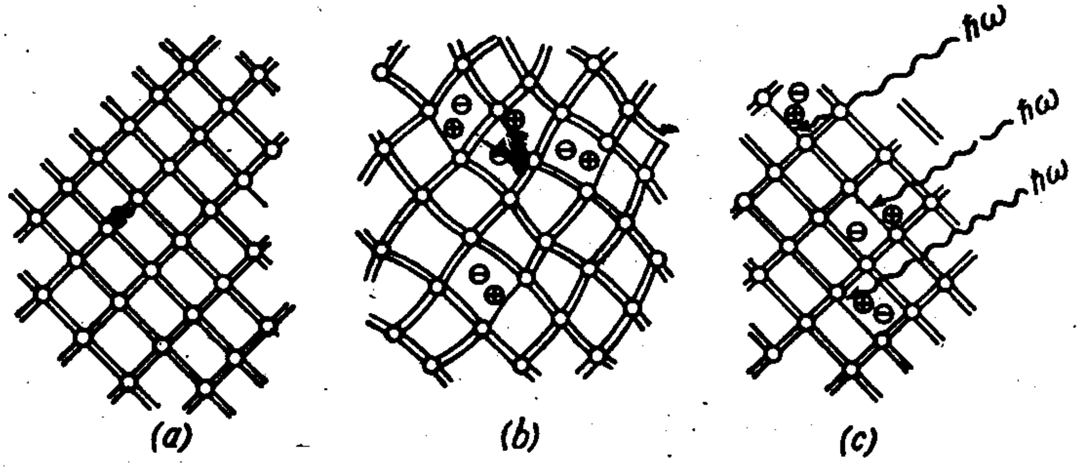


Fig. 7. The appearance of electrical conductivity due to thermal lattice vibrations and to the irradiation of the semiconductor

Let us discuss the model of conduction in semiconductors using silicon as an example.

The silicon atom has 14 electrons which are divided among the electron shells in the following way: $(1s^2)$ $(2s^2)$ $(2p^6)$ $(3s^2)$ $3p^2$.

The incomplete outer shell contains four electrons. The electron cloud of an atom interacting to form a crystal is of a tetrahedral pattern. Such an electron cloud pattern determines the lattice type of the silicon crystal—the diamond type. The diamond-type lattice (Fig. 6) is cubic. Every atom is bound to four nearest silicon atoms by covalent bonds, and all of its four valence electrons take part in these bonds. In the ideal lattice all electrons are bound, there are no free carriers, and, correspondingly, the application of an electric field does not result in an electric current. To produce current some electrons must be freed from their bond. To tear an electron away from its bond energy must be expended. This energy may be supplied to the crystal in the form of the energy of photons, or of particles, or in the form of the energy of thermal lattice vibrations. At room temperature the energy needed to free one electron in silicon is 1.08 eV. Note that the freeing of one electron produces one incomplete bond. Figure 7 shows schematic diagrams of the ideal crystal (*a*), of a crystal with a free electron produced by thermal vibrations of the lattice (*b*) and of a crystal with a free electron produced as a result of a photon with appropriate energy being absorbed (*c*). The freed electron will roam in the crystal at random. If electron approaches an atom with incomplete bonds, it can join the atom. transmitting its energy to the lattice or emitting a light quantum. *The process of bonding a free electron is termed recombination. It is the opposite of the process of freeing a bonded electron (generation of a free electron).* Note that recombination results not only in the disappearance of one free electron but of one incomplete bond as well. The numbers of free electrons and vacant (incomplete) bonds are equal. In the vicinity of an incomplete bond there is an excess positive charge not compensated by the electron charge. However, the total charge of a sufficiently large volume of the substance is, as before, zero. Thus the appearance of electrons and vacant bonds in equal numbers does not change the electric neutrality of the crystal as a whole. If electric field **E** is applied to the crystal, free electrons taking part in random thermal motion will be acted upon by the force $e_n\mathbf{E}$ and will begin to drift against the field. If we denote electron concentration by n, their mobility by μ_n, we will be able to write for the electron current density

$$\mathbf{j}_n = e_n \mu_n n \mathbf{E} = \sigma_n \mathbf{E}. \qquad (5.1)$$

Here e_n is the electron charge.

In semiconductors the conductivity actually depends on the ambient since by changing the intensity of illumination, radiation or temperature the charge carrier concentration can be widely varied while in metals the number of electrons remains independent of the ambient. This, however, is not the only difference between metals and semiconductors. The latter exhibit conductivities of two types.

The incomplete bond can, in fact, move from atom to atom as a result of electron motion, i.e. it can wander at random in the crystal. When external electric field **E** is applied the bound electrons will be acted upon by the force $e_n\mathbf{E}$ and will be moving against the field and occupy vacant bonds.

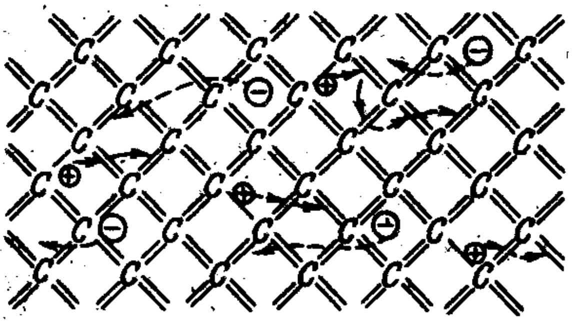


Fig. 8. The motion of electrons and holes in the electric field

The motion of bound electrons in the ideal crystal in which all the bonds are occupied is excluded by the Pauli principle. The availability of vacant bonds enables valence electrons to move against the field. This means that their mobility depends on the number of vacant bonds usually called holes. In this way the mass of the valence electrons, too, contributes to semiconductor conductivity.

If the concentration of bound electrons is N and their mobility μ_N, their current density is

$$\mathbf{j}_N = e_n\mu_N N\mathbf{E}. \qquad (5.2)$$

Thus, in semiconductors there are two kinds of charge carriers: free electrons and bound electrons. Therefore

$$\mathbf{j} = \mathbf{j}_n + \mathbf{j}_N = (e_n\mu_n n + e_n\mu_N N)\,\mathbf{E}. \qquad (5.3)$$

It is, however, more convenient to consider not the motion of the mass of valence (bound) electrons but the motion of vacant bonds (holes). *The motion of (collectivized) electron along the bonds against the field* (Fig. 8) *results in the motion of the hole in the direction of the field*; this motion being equivalent to the motion of a positive charge e^+ in the direction of the field. It must, however, be remembered that *the positive charge moves in the direction of the field not because it is acted upon by the force* $e^+\mathbf{E}$ *but because the electrons move against the field.*

To move a positive charge e^+ directly an ionized silicon atom should be moved. Since, however, the mobility of ions is many orders of magnitude less than the mobility of electrons, ionic current is practically absent in silicon crystals. For this reason the only cause of the motion of uncompensated nuclear charge is the motion of the valence electrons. Denoting the hole concentration by p and their mobility by μ_p, we can write the current density of the mass of bound electrons in the form

$$\mathbf{j}_N = e_N\mu_N N\mathbf{E} = e_p\mu_p p\mathbf{E} = \mathbf{j}_p. \qquad (5.4)$$

Here μ_p should not depend on p. This will be shown below with the help of the Brillouin zone method.

The mechanism of conduction by bound electrons was termed hole conduction. Holes with the charge $e_p = e^+$ are regarded as quasi-particles the motion of which is quite equivalent to the motion of valence electrons. To describe physical phenomena with consistent results using the hole concept all their properties must be correctly assessed. Only then will it be possible to discuss the motion of holes without discussing the motion of the mass of valence electrons. This will be done in Sec. 25.

6. INTRINSIC AND EXTRINSIC CONDUCTIVITIES

A semiconductor with equal concentrations of electrons and holes $n = p$ is termed intrinsic. In it

$$\mathbf{j} = \sigma\mathbf{E} = (e_n\mu_n n + e_p\mu_p p)\,\mathbf{E} = (\sigma_n + \sigma_p)\,\mathbf{E}. \qquad (6.1)$$

Using b for the ratio of the mobility moduli (assuming mobility to be a scalar)

$$b = \frac{|\mu_n|}{|\mu_p|} \qquad (6.2)$$

we may write for the conductivity of an intrinsic semiconductor

$$\sigma = e_n\mu_n n + e_p\mu_p p = e_p p\mu_p\,(1 + b) \qquad (6.3)$$

However, for device operation the concentrations of electrons and holes should be different. This difference is produced by introducing impurities, also called doping. This leads to *extrinsic (impurity) conductivity*. To understand the essence of extrinsic conductivity let us consider silicon doped with elements of the fifth and the third groups. Suppose an arsenic atom replaces a silicon atom in the crystal lattice. The outer shell of the arsenic atom has five electrons. Four of them will take part in forming covalent bonds with nearest silicon atoms. The fifth electron is unable to take part in bonding since all the bonds are filled. At the same time it is acted upon by the neighbouring silicon atoms, hence the decrease of its bonding energy to the arsenic atom Calculations show that this energy decreases about ε^2 times (ε is the relative dielectric permeability: for silicon $\varepsilon = 12$). This means that the fifth electron can be freed at the expense of energy over a hundred times smaller than the one needed to tear away an electron from the matrix atom (in this case—the silicon atom). For this reason impurities are easily ionized, and a large number of free electrons appear in the crystal—much larger than in pure silicon. We see from here that doping silicon with group V elements results in a much larger concentration of free electrons than in pure silicon. *Impurities which supply electrons are termed donors.* Impurity atoms having surrendered their electrons become charged ions (As^+). But since positive impurity ions cannot play an appreciable part in electric current, the donor impurity acts only as a source of free electrons.

Ionization of lattice atoms takes place not only in a doped semiconductor but in an undoped one too; however, the concentration of free carriers in this case will be much smaller than in the former. For this reason the concentration of electrons in a donor-doped semiconductor will be much greater than the concentration of holes. *Current in such a crystal is carried mainly by electrons by reason of which they are termed majority carriers, while the holes are termed minority carriers. Such a semiconductor bears the name of electron, or n-type, semiconductor.* The conductivity of an electron-type semiconductor may be written in the form

$$\sigma = \sigma_n + \sigma_p \cong \sigma_n = e_n \mu_n n, \tag{6.4}$$

since $p \ll n$ and $\sigma_p \ll \sigma_n$.

The conductivity type can in some cases be determined with the help of a simple valence rule: if the valence of the impurity atom is a unit larger than that of the matrix atom, the impurity is a donor. Thus, group V elements are donors in silicon and germanium. In $A^{III}B^{V}$ compounds group VI elements act as donors. This rule is, however, far from being universal.

Consider now another case. Suppose an atom of indium, or of any other group III element, is introduced into silicon. The indium atom has three valence electrons by reason of which one of its bonds with silicon atoms will remain incomplete. To fill this bond it is necessary to transfer to the indium atom one electron belonging to some silicon atom. The energy needed to transfer an electron from a matrix atom to an impurity atom which thereby becomes a negatively charged ion is much smaller than the one needed to free the electron. Therefore the transitions of electrons from the matrix atoms to impurity atoms will occur at a much higher rate than the acts of ionization of the matrix atoms leading to the creation of free electrons. The negative charge is localized at the impurity atom site and therefore plays no part in the current. The incomplete bond (a hole in the bonds) between two matrix atoms carries a positive charge.

An impurity atom which accepts an electron is termed acceptor. In a semiconductor doped with acceptors the hole concentration greatly surpasses the free electron concentration, and because of this its conductivity is mainly of the hole type:

$$\sigma = \sigma_n + \sigma_p \cong \sigma_p = e_p \mu_p p \qquad (6.5)$$

since $n \ll p$ and $\sigma_n \ll \sigma_p$.

In this case holes are majority carriers and the electrons—minority carriers. An acceptor-doped semiconductor is termed hole, or p-type semiconductor. Often acceptors are elements the valence of which is a unit less than the valence of the matrix atoms, for instance, group III elements in Si, Ge; group II elements in $A^{III}B^{V}$ compounds.

Since carrier concentration in an impurity semiconductor exceeds carrier concentration in the intrinsic semiconductor of the same kind, the resistance of such a semiconductor doped with impurity of one type is inferior to the resistance of the pure semiconductor. In other words, *doping of a semiconductor decreases its resistance.*

In case both types of impurity—donors and acceptors—are present in a semiconductor compensation takes place. When donor and acceptor concentrations are equal (provided both are fully ionized) a doped semiconductor behaves like the intrinsic one, its resistance is high. Such semiconductors are termed *compensated.* Semiconductors the conductivity of which depends on the type of impurity are termed amphoteric.

Summary of Secs. 4-6

1. The class of semiconductors includes materials the conductivity of which largely depends on their composition and structure, and on the ambient. As a rule, the conductivity of semiconductors increases when energy is supplied to them in the form of heat, illumination, and nuclear radiation. Conductivity depends also on pressure and on external electric and magnetic fields.

2. There are two conduction mechanisms in semiconductors: either free electrons or holes can act as charge carriers. This is termed bipolar conductivity. Hole conductivity is the result of motion of bound electrons along the bonds.

3. In a pure semiconductor the numbers of holes and electrons are equal. Such a semiconductor is termed intrinsic. The impurity which supplies free electrons is termed a donor, the one which supplies holes is termed an acceptor. Charge carriers present in the semiconductor in higher concentration are termed majority carriers, those in lesser concentration are termed minority carriers.

A semiconductor with equal concentrations of acceptor- and donor-type impurities is termed compensated.

Chapter II

THE FUNDAMENTALS OF THE BAND THEORY OF SEMICONDUCTORS

7. THE SCHRÖDINGER EQUATION FOR THE CRYSTAL

The crystal is a unified system of light (electrons) and heavy (nuclei) particles. Since the electrons of the inner atomic shells do not take part in electro-physical processes in the crystal, we shall consider the system as consisting of the electrons of the outer shells and of ions.

Denote the electron co-ordinates by $\mathbf{r}_1$, $\mathbf{r}_2$, ... and the ionic coordinates by $\mathbf{R}_1$, $\mathbf{R}_2$, ... The stationary state of the particles is described by the Schrödinger equation:

$$\hat{\mathbf{H}}\Psi = E\Psi \qquad (7.1)$$

where $\hat{\mathbf{H}}$ is the crystal Hamiltonian, Ψ, its wave function, E, its eigenvalue, or the crystal energy. The crystal wave function is a function of co-ordinates of all the particles:

$$\Psi = \Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots; \mathbf{R}_1, \mathbf{R}_2, \ldots) = \Psi(\mathbf{r}_i; \mathbf{R}_\alpha), \qquad (7.2)$$

where, for the sake of brevity, the electron co-ordinates are denoted by $\mathbf{r}_i$, and the ionic co-ordinates, by $\mathbf{R}_\alpha$.

The Hamilton operator includes all types of energy, i.e. (1) *the kinetic energy of the electrons* $\hat{\mathbf{T}}_e$:

$$\hat{\mathbf{T}}_e = \sum_i \hat{\mathbf{T}}_i = \sum_i \left(-\frac{\hbar^2}{2m}\Delta_i\right), \qquad (7.3)$$

where m is the electron mass, $\hbar = \frac{h}{2\pi}$ is the Planck constant h divided by 2π, $\Delta_i = \nabla_i^2$ is the Laplace operator for the i-th electron

$$\Delta_i = \nabla_i^2 = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2}; \qquad (7.4)$$

(2) *the kinetic energy of the ions* $\hat{\mathbf{T}}_Z$:

$$\hat{\mathbf{T}}_Z = \sum_\alpha \hat{\mathbf{T}}_\alpha = \sum_\alpha \left(-\frac{\hbar^2}{2M_\alpha}\Delta_\alpha\right), \qquad (7.5)$$

where M_α is the ionic mass, and

$$\Delta_\alpha = \frac{\partial^2}{\partial X^2} + \frac{\partial^2}{\partial Y^2} + \frac{\partial^2}{\partial Z^2}; \tag{7.6}$$

(3) *the energy of electron-electron interaction* $\hat{U}_e$:

$$\hat{U}_e = \frac{1}{2} \sum_{i,\, j \neq i} \frac{e^2}{4\pi\varepsilon_0 |\mathbf{r}_i - \mathbf{r}_j|} = \frac{1}{2} \sum_{i \neq j} \hat{U}_{ij}; \tag{7.7}$$

(4) *the energy of ion-ion interaction* $\hat{U}_Z$:

$$\hat{U}_Z = \frac{1}{2} \sum_{\alpha \neq \beta} \frac{Z_\alpha Z_\beta e^2}{4\pi\varepsilon_0 |\mathbf{R}_\alpha - \mathbf{R}_\beta|} = \frac{1}{2} \sum_{\alpha \neq \beta} \hat{U}_{\alpha\beta}, \tag{7.8}$$

where $Z_\alpha e$, $Z_\beta e$ are the charges of α and β ions;

(5) *the energy of electron-ion interaction* $\hat{U}_{eZ}$:

$$\hat{U}_{eZ} = -\sum_{i,\, \alpha} \frac{Z_\alpha e^2}{4\pi\varepsilon_0 |\mathbf{r}_i - \mathbf{R}_\alpha|} = \sum_{i,\, \alpha} \hat{U}_{i\alpha}; \tag{7.9}$$

(6) *the energy of all the particles in an external field* $\hat{V}$:

$$\hat{V} = \hat{V}(\mathbf{r}_1, \mathbf{r}_2, \ldots; \mathbf{R}_1, \mathbf{R}_2, \ldots). \tag{7.10}$$

Correspondingly, the Hamiltonian of a crystal in an external field may be written in the form:

$$\hat{H} = \hat{T}_e + \hat{T}_Z + \hat{U}_e + \hat{U}_Z + \hat{U}_{eZ} + \hat{V}; \tag{7.11}$$
$$\hat{H}\Psi = E\Psi.$$

The Schrödinger equation (7.11) contains $3(Z+1)N$ variables, where N is the number of atoms in the crystal. Since 1 cm^3 of a crystal contains about $5 \cdot 10^{22}$ atoms, the number of variables for $Z=14$ reaches $2 \cdot 10^{24}$ cm^{-3}. Obviously, this equation cannot be solved in a general form. This is not only because of involved calculations but mainly because of difficulties of a principal nature, for modern quantum mechanics lacks the means of solving many-particle problems. To solve the Schrödinger equation for a system of interacting particles the problem must be transformed into a problem for a system of non-interacting particles. In the latter case the equation for a system of particles may be broken up into a set of equations each of which describes the motion of a single particle. To wit if the system Hamiltonian may be represented by a sum of Hamiltonians:

$$\hat{H} = \sum_k \hat{H}_k, \tag{7.12}$$

where each $\hat{H}_k$ is the function of the co-ordinates of only one kth particle,

$$\hat{H}_k = -\frac{\hbar^2}{2m_k}\Delta_k + \hat{U}_k(\mathbf{r}_k), \tag{7.13}$$

i.e. the particles do not interact, then the Schrödinger equation may be solved in the following way. *The system wave function is represented by the product of wave functions which describe individual particles, and the energy of the system is equated to the sum of particle energies.*

$$\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots) = \psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2)\ \ldots \tag{7.14}$$

and

$$E = \sum_k E_k, \tag{7.15}$$

where E_k and ψ_k are related by the expression

$$\hat{H}_k\psi_k(\mathbf{r}_k) = E_k\psi_k(\mathbf{r}_k) \tag{7.16}$$

which is readily proved by simply substituting (7.14) into the Schrödinger equation for a particle system, whence with the aid of (7.16) one obtains (7.15).

The transition from the system (7.11) of interacting particles to a system of non-interacting particles is possible only as a result of an approximate solution of the equation subject to several more or less obvious simplifications. Below we will presume the external fields to be absent,

$$\hat{V}(\mathbf{r}_1, \ldots; \mathbf{R}_1, \ldots) = 0. \tag{7.17}$$

Prior to simplifying the Schrödinger equation we will write down the expression for the energy of the crystal

$$E = \int \Psi^*(\mathbf{r}_1, \ldots; \mathbf{R}_1, \ldots)\,\hat{H}\Psi(\mathbf{r}_1, \ldots; \mathbf{R}_1, \ldots)\,d\tau, \tag{7.18}$$

where integration is performed over the co-ordinates of all the particles:

$$d\tau = dx_1\,dy_1\,dz_1\ldots dX_1\,dY_1\,dZ_1\ldots = d\tau_e\,d\tau_Z. \tag{7.19}$$

The wave function $\Psi(\mathbf{r}_1, \ldots; \mathbf{R}_1, \ldots)$ enables the motion of any particle belonging to the crystal to be found. Considering the minimum energy states it would have been possible to obtain theoretically the crystal structure and its probable modifications.

8. THE ADIABATIC APPROXIMATION

The adiabatic approximation, or the Born-Oppenheimer approximation, takes account of the difference in motion of light (electrons) and heavy (ions) particles. Important for fast moving electrons is the instantaneous position of the ions. At the same time the motion of ions is affected not by instantaneous position of the electrons but only by their averaged motion. This statement will be readily understood as applied to many-electron atoms. Obviously, the ion, on account of its mass, does not follow in its motion the motion of every electron but moves only in the averaged field of all the electrons. At the same time the ion moving relatively slowly is accompanied by the electrons, and the atom thus remains intact. The same should hold for a crystal.

The roughest assumption is that the ions remain stationary: $\mathbf{R}_\alpha = \mathbf{R}_\alpha^0$. In this case the Schrödinger equation is greatly simplified. In fact, the kinetic energy of the ions in this case turns zero, and the energy U_Z of interaction of the ions becomes a constant which can also be made zero by an appropriate choice of the reference energy. *Let us write the expression for the Hamiltonian, now termed the electron Hamiltonian and denoted by $\hat{\mathbf{H}}_e$, taking into account that $\hat{\mathbf{T}}_Z = 0$ and $\hat{\mathbf{U}}_Z = 0$:*

$$\hat{\mathbf{H}}_e = \hat{\mathbf{T}}_e + \hat{\mathbf{U}}_e + \hat{\mathbf{U}}_{eZ}. \tag{8.1}$$

The electron wave function will be denoted by Ψ_e. It should depend on the electron co-ordinates $\mathbf{r}_i$ and the stationary ionic co-ordinates $\mathbf{R}_\alpha^0$. $\Psi_e(\mathbf{r}_i; \mathbf{R}_\alpha^0)$, when integrated over the electron co-ordinates, should be normalized for all values of the ionic co-ordinates:

$$\int \Psi_e^* (\mathbf{r}_1, \ldots; \mathbf{R}_1^0, \ldots) \Psi_e (\mathbf{r}_1, \ldots; \mathbf{R}_1^0, \ldots) d\tau_e. \tag{8.2}$$

The Schrödinger equation may be written in the following form:

$$\hat{\mathbf{H}}_e \Psi_e = E_e \Psi_e; \tag{8.3}$$

$$\left[\sum_i \left(-\frac{\hbar^2}{2m} \Delta_i\right) + \frac{1}{8\pi\varepsilon_0} \sum_{i \neq j} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} - \frac{1}{4\pi\varepsilon_0} \sum_{i,\,\alpha} \frac{Ze^2}{|\mathbf{r}_i - \mathbf{R}_\alpha^0|}\right] \Psi_e = E_e \Psi_e.$$

$\mathbf{R}_\alpha^0$ now enters the equation not as a variable of the differential equation, but as a parameter the choice of which ultimately affects the wave function and the crystal energy E_e:

$$E_e = \int \Psi_e^* \hat{\mathbf{H}}_e \Psi_e d\tau_e = E_e (\mathbf{R}_1^0, \mathbf{R}_2^0, \ldots), \tag{8.4}$$

E_e is the energy of electrons moving in the field of stationary ions.

The assumption of stationary ions is, however, too rough. We may assume the ions to be in motion and take account of this fact by introducing an atomic wave function $\Phi_Z(\mathbf{R}_1, \ldots)$. With this in mind we will write out the crystal Hamiltonian in the following form. $\hat{\mathbf{H}}_z$ is the *operator*

$$\hat{\mathbf{H}}_Z=\sum_{\alpha}\left(-\frac{\hbar^2}{2M_\alpha}\Delta_\alpha\right)+\hat{\mathbf{U}}_Z+\hat{E}_e(\ldots, \mathbf{R}_\alpha, \ldots). \qquad (8.5)$$

which we will term the *ionic part of the crystal Hamiltonian*. Evidently, the crystal Hamiltonian may be expressed in terms of $\hat{\mathbf{H}}_e$ and $\hat{\mathbf{H}}_Z$. Taking into account (7.11), (7.17), (8.1), and (8.5), we write

$$\hat{\mathbf{H}}=\hat{\mathbf{H}}_e+\hat{\mathbf{H}}_Z-\hat{E}_e. \qquad (8.6)$$

Let us write the crystal wave function Ψ as a product

$$\Psi(\ldots, \mathbf{r}_i, \ldots; \ldots, \mathbf{R}_\alpha, \ldots)=$$
$$=\Psi_e(\ldots, \mathbf{r}_i, \ldots; \ldots, \mathbf{R}_\alpha, \ldots)\Phi_Z(\ldots, \mathbf{R}_\alpha, \ldots), \qquad (8.7)$$

substitute it into the Schrödinger equation for the crystal and find the equation for $\Phi_Z(\mathbf{R}_1, \ldots)$,

$$\hat{\mathbf{H}}\Psi=(\hat{\mathbf{H}}_e+\hat{\mathbf{H}}_Z-\hat{E}_e)\Psi_e\Phi_Z=$$
$$=\Phi_Z\hat{\mathbf{H}}_e\Psi_e+\hat{\mathbf{H}}_Z\Psi_e\Phi_Z-\hat{E}_e\Psi_e\Phi_Z=E\Psi_e\Phi_Z. \qquad (8.8)$$

Or we may, taking into account (8.3), write

$$\hat{\mathbf{H}}\Psi=\hat{\mathbf{H}}_Z\Psi_e\Phi_Z=E\Psi_e\Phi_Z=E\Psi. \qquad (8.9)$$

Since Ψ_e is dependent on ionic co ordinates, $\hat{\mathbf{H}}_Z$ will affect Ψ_e. Let us find, for instance, $\Delta_\alpha\Psi_e\Phi_Z$:

$$\Delta_\alpha\Phi_Z\Psi_e=\nabla_\alpha(\Psi_e\nabla_\alpha\Phi_Z+\Phi_Z\nabla_\alpha\Psi_e)=$$
$$=\Psi_e\Delta_\alpha\Phi_Z+\Phi_Z\Delta_\alpha\Psi_e+2(\nabla_\alpha\Phi_Z\cdot\nabla_\alpha\Psi_e).$$

With reference to this formula let us rewrite (8.9)

$$\hat{\mathbf{H}}_Z\Psi_e\Phi_Z=\sum_{\alpha}\left(-\frac{\hbar^2}{2M_\alpha}\right)[\Psi_e\Delta_\alpha\Phi_Z+\Phi_Z\Delta_\alpha\Psi_e+2(\nabla_\alpha\Psi_e\cdot\nabla_\alpha\Phi_Z)]+$$
$$+U_Z\Psi_e\Phi_Z+E_e\Psi_e\Phi_Z=E\Psi_e\Phi_Z. \qquad (8.10)$$

Next we premultiply (8.10) by Ψ^* and integrate it over electron co-ordinates. Taking into account that

$$\int\Psi_e^*\Psi_e d\tau_e=1 \qquad (8.11)$$

we obtain

$$\sum_{\alpha}\left(-\frac{\hbar^2}{2M_\alpha}\right)\Big[\Delta_\alpha\Phi_Z+\Phi_Z\int\Psi_e^*\Delta_\alpha\Psi_e\,d\tau_e+ +2\Big(\nabla_\alpha\Phi_Z\cdot\int\Psi_e^*\nabla_\alpha\Psi_e\,d\tau_e\Big)\Big]+U_Z\Phi_Z+E_e\Phi_Z=E\Phi_Z. \qquad (8.12)$$

Recalling the notation (8.5) for $\hat{H}_Z$, we may rewrite (8.12) in the following form

$$\hat{H}_Z\Phi_Z=E\Phi_Z+\sum_{\alpha}\frac{\hbar^2}{2M_\alpha}\Big[\Phi_Z\int\Psi_e^*\Delta_\alpha\Psi_e\,d\tau_e+ +2\Big(\nabla_\alpha\Phi_Z\cdot\int\Psi_e^*\nabla_\alpha\Psi_e\,d\tau_e\Big)\Big]. \qquad (8.13)$$

Premultiplying the equation (8.13) by Φ_Z^* and integrating over ionic co-ordinates, we obtain the expression for the energy of the crystal

$$\int\Phi_Z^*\hat{H}_Z\Phi_Z\,d\tau_Z=E+\delta E. \qquad (8.14)$$

To assess the quantity

$$\delta E=\sum_{\alpha}\frac{\hbar^2}{2M_\alpha}\int\Psi_e^*\Delta_\alpha\Psi_e\,d\tau_e+ +\sum_{\alpha}\frac{\hbar^2}{2M_\alpha}\Big(2\int\Phi_Z^*\nabla_\alpha\Phi_Z\,d\tau_Z\cdot\int\Psi_e^*\nabla_\alpha\Psi_e\,d\tau_e\Big), \qquad (8.15)$$

it is necessary to make certain assumptions concerning the nature of the dependence of Ψ_e on the co-ordinates $\mathbf{r}_l$ and $\mathbf{R}_\alpha$. If the electron interaction is neglected in $\hat{H}_e$, it turns into a Hamiltonian of a system of non-interacting particles. The wave function of such a system will consist of a combination of wave functions which depend on the difference $|\mathbf{r}_l-\mathbf{R}_\alpha|$. But in this case

$$\nabla_\alpha^n\Psi_e=(-1)^n\nabla_l^n\Psi_e, \qquad (8.16)$$

and the first term of (8.15) may be rewritten in the form

$$\sum_{\alpha}\frac{\hbar^2}{2M_\alpha}\int\Psi_e^*\Delta_\alpha\Psi_e\,d\tau_e=\sum_{l,\,\alpha}\frac{\hbar^2}{2M_\alpha}\int\Psi_e^*\Delta_l\Psi_e\,d\tau_e= =-\sum_{\alpha,\,l}\frac{m}{M_\alpha}\int\Psi_e^*\left(-\frac{\hbar^2}{2m}\Delta_l\right)\Psi_e\,d\tau_e=-\sum_{l,\,\alpha}\frac{m}{M_\alpha}\langle T_l\rangle, \qquad (8.17)$$

where $\langle T_l\rangle$ is the mean kinetic energy per electron. Hence

$$\sum_{\alpha}\frac{\hbar^2}{2M_\alpha}\int\Psi_e^*\Delta_\alpha\Psi_e\,d\tau_e=-ZN\cdot\frac{m}{M}\langle T_l\rangle. \qquad (8.18)$$

But this term may be neglected in comparison with E; the *error introduced thereby into the expression for the crystal energy being quite small—of the order of the ratio of the electron mass to the ionic mass* which, for example, for Ge is about 10^{-5}.

The second term in (8.15) may be assessed in the same way:

$$\frac{\hbar^2}{M_\alpha}\left(\int \Phi_Z^* \nabla_\alpha \Phi_Z \, d\tau_Z \cdot \int \Psi_e^* \nabla_\alpha \Psi_e \, d\tau_e\right) = -\frac{1}{M_\alpha}(\langle \mathbf{p}_\alpha \rangle \cdot \langle \mathbf{p}_i \rangle). \qquad (8.19)$$

Since in classical statistics we have for a system in a state of thermodynamical equilibrium

$$\left\langle \frac{\mathbf{p}_i^2}{2m} \right\rangle = \left\langle \frac{\mathbf{p}_\alpha^2}{2M_\alpha} \right\rangle \qquad (8.20)$$

and

$$\langle p_e \rangle^2 = \frac{8}{3\pi} \langle p_e^2 \rangle, \qquad (8.21)$$

it follows that

$$\langle p_e \rangle \cong \sqrt{\frac{m}{M}} \langle p_Z \rangle \qquad (8.22)$$

and *the second term in* (8.15) *is of the order of* $\sqrt{\frac{m}{M}}$ *times the total crystal energy.*

We chose the most "disadvantageous" case when the electron wave function is a combination of atomic wave functions (the so-called strongly bound electron approximation). If the electron wave function is independent of ionic co-ordinates (the free electron approximation) *both corrections in* (8.15) *will be zero.*

In this way we arrive at the conclusion that by neglecting both corrections we introduce an error in the value of E not greater than $\sqrt{\frac{m}{M}} E$ (for germanium $\sqrt{\frac{m}{M}}$ is about 0.3%). Hence, to determine the crystal energy with sufficient accuracy one can solve the equation

$$\hat{\mathrm{H}}_Z \Phi_Z = E_Z \Phi_Z = E \Phi_Z. \qquad (8.23)$$

In other words, *total crystal energy with great accuracy coincides with the eigenvalue of the ionic part of the Hamiltonian.*

We see from here that the adiabatic approximation for a crystal with the Hamiltonian (7.11) gives a sufficiently accurate energy value if the assumption is made that the crystal wave function may be written in the form

$$\Psi(\mathbf{r}_1, \ldots; \mathbf{R}_1, \ldots) = \Phi_Z(\mathbf{R}_1, \ldots) \Psi_e(\mathbf{r}_1, \ldots; \mathbf{R}_1, \ldots), \qquad (8.24)$$

Ψ_e and Φ_Z being determined from the equations

$$\hat{H}_e\Psi_e = E_e\Psi_e, \tag{8.25}$$

$$\hat{H}_Z\Phi_Z = E_Z\Phi_Z = E\Phi_Z. \tag{8.26}$$

Thus, *in the adiabatic approximation the electron wave function is determined by the instantaneous position of the ions* (the term U_{eZ} in $\hat{H}_e$), *the ionic wave function, on the other hand, being determined by the averaged electron field* (the term E_e in $\hat{H}_Z$).

Summary of Secs. 7-8

1. The crystal Hamiltonian accounts for all kinds of electron and ion energy: for the kinetic energy of the electrons and ions, for the interaction energy of the particles and for the energy of all particles in external fields. The Schrödinger equation for a crystal takes the form (7.11). It contains $3(Z+1)N$ variables and cannot be solved in a general form.

2. The adiabatic approximation enables the co-ordinates of the electrons and ions to be separated and thereby leads to simplification of the equation (7.11).

The crystal Hamiltonian is subdivided into the electron $\hat{H}_e$ (8.1) and the atomic $\hat{H}_Z$ (8.6) parts, and the crystal wave function Ψ is represented by the product of the ion Φ_Z and the electron Ψ_e wave functions:

$$\Psi = \Phi_Z\Psi_e. \tag{8.1s}$$

Simultaneously

$$\hat{H}_e\Psi_e = E_e\Psi_e, \tag{8.2s}$$

$$\hat{H}_Z\Phi_Z = E\Phi_Z, \tag{8.3s}$$

and the terms

$$\sum_\alpha \frac{\hbar^2}{2M_\alpha}\left[\int \Psi_e^*\Delta_\alpha\Psi_e\,d\tau_e + 2\left(\int \Phi_Z^*\nabla_\alpha\Phi_Z\,d\tau_Z\cdot\int\Psi_e^*\nabla_\alpha\Psi_e\,d\tau_e\right)\right] \tag{8.4s}$$

may be disregarded as negligible. The accuracy in energy reaches the order of $\sqrt{\frac{m}{M}}$, but some processes relating to thermal lattice vibrations fall outside the scope of theoretical consideration. The ionic wave function Φ_Z is determined by the averaged motion of the electrons (the term E_e of $\hat{H}_Z$), while Ψ_e depends on the instantaneous position of the ions (the term U_{eZ} in $\hat{H}_e$). Neglecting the correction δE means that the action of Δ_α on Ψ is reduced only to the action on Φ_Z. This reflects the fact that ions move more slowly than electrons.

9. SINGLE ELECTRON APPROXIMATION

As a result of adiabatic approximation the electron wave function should satisfy the equation

$$\hat{H}_e\Psi_e = E_e\Psi_e \tag{9.1}$$

or

$$\left[\sum_i\left(-\frac{\hbar^2}{2m}\Delta i\right)+\frac{1}{2}\sum_{i\neq j}U_{ij}+\sum_{i,\alpha}U_{i\alpha}\right]\Psi_e = E_e\Psi_e. \tag{9.2}$$

This equation, too, cannot be solved. It should first be transformed into an equation for one particle. It follows directly from (9.2) that this equation for an electron system turns into a system of equations if it is assumed that the electrons do not interact ($U_{ij}=0$). Therefore the problem arises how to take account of the electron interaction so that ultimately we would be able to deal with a system of non-interacting instead of interacting electrons. This is achieved through the introduction of the so-called self-consistent electron field. Let us take some ith electron. It is acted upon by the field of all ions and remaining electrons. Suppose we are able with the aid of an external source to establish, every instant of time, the same field at the point of location of the ith electron as is established by other electrons. Denote the potential energy of the ith electron in this field by Ω_i. It will obviously depend only on the co-ordinates of the ith electron: $\Omega_i=\Omega_i\ (\mathbf{r}_i)$.

If we were able to construct such fields for every electron, we could substitute the sum of $\Omega_i\ (\mathbf{r}_i)$ terms for the energy of electron-electron interaction:

$$\frac{1}{2}\sum_{i\neq j}{}'\frac{e^2}{4\pi\varepsilon_0|\mathbf{r}_i-\mathbf{r}_j|}\to\sum_i\Omega_i(\mathbf{r}_i). \tag{9.3}$$

The potential energy $\Omega_i\ (\mathbf{r}_i)$ of the ith electron depends not only on the motion of every other electron but it also indirectly depends on the motion of the ith electron itself, since this motion affects the motion of other electrons. Therefore the field $\Omega_i\ (\mathbf{r}_i)$ not only determines the motion of the ith electron but is itself dependent upon its motion, and for this reason is termed self-consistent. In principle, the field $\Omega_i\ (\mathbf{r}_i)$ can be found with the aid of the method of successive approximations. We will discuss below the idea of a method for determining this field which was originally developed for the atom; first we will consider a consequence of the introduction of the self-consistent field. Supposing that such a field was determined, we can write the Hamiltonian $\hat{H}_e$ in the form

$$\hat{H}_e=\sum_i\left(-\frac{\hbar^2}{2m}\Delta_i\right)+\frac{1}{2}\sum_{i\neq j}U_{ij}+\sum_{i,\alpha}U_{i\alpha}=$$
$$=\sum_i\left(-\frac{\hbar^2}{2m}\Delta_i\right)+\sum_i\Omega_i(\mathbf{r}_i)+\sum_i\left(\sum_\alpha U_{i\alpha}\right)=\sum_i\hat{H}_i, \tag{9.4}$$

where the Hamiltonian for the ith electron is

$$\hat{H}_i=-\frac{\hbar^2}{2m}\Delta_i+\Omega_i(\mathbf{r}_i)+U_i(\mathbf{r}_i); \tag{9.5}$$

$\Omega_i(\mathbf{r}_i)$ is the potential energy of the ith electron in the field of all other electrons and $U_i(\mathbf{r}_i)$, in the field of all ions.

Since now the Hamiltonian contains no electron interaction energy, the wave function of the electron system assumes the form of a product of the wave functions of separate electrons, and the energy of the system becomes equal to the sum of energies of separate electrons:

$$\Psi_e(\mathbf{r}_1, \mathbf{r}_2, \ldots)=\prod_i\psi_i(\mathbf{r}_i), \tag{9.6}$$

$$E_e=\sum_i E_i, \tag{9.7}$$

and

$$\hat{H}_i\psi_i=E_i\psi_i, \tag{9.8}$$

i.e. *the introduction of the self-consistent field reduces the many-electron problem to the one-electron problem.*

To assess the form of $\Omega_i(\mathbf{r}_i)$ let us write out the Schrödinger equation for the electron part of the crystal Hamiltonian in two ways:

$$\hat{H}_e\Psi_e=\left[\sum_i\left(-\frac{\hbar^2}{2m}\Delta_i\right)\Psi_e+\frac{1}{2}\sum_{i\neq j}U_{ij}\Psi_e+\sum_i U_i(\mathbf{r}_i)\Psi_e\right]=E_e\Psi_e \tag{9.9}$$

and

$$\hat{H}_e\Psi_e=\left[\sum_i\left(-\frac{\hbar^2}{2m}\Delta_i\right)\Psi_e+\sum_i\Omega_i(\mathbf{r}_i)\Psi_e+\sum_i U_i(\mathbf{r}_i)\Psi_e\right]=E_e\Psi_e. \tag{9.10}$$

These two forms of the same equation give the operator $\Omega_i(\mathbf{r}_i)$ which, however, cannot be determined with the aid of the equation

$$\Omega_i(\mathbf{r}_i)=\frac{1}{2}\sum_{j(j\neq i)}U_{ij} \tag{9.11}$$

since $\Omega_i(\mathbf{r}_i)$ should depend only on the co-ordinates of the ith electron; $\sum_{j(i\neq j)} U_{ij}$, on the other hand, depends on the co-ordinates of all electrons. Obviously, the equations (9.11) and (9.3) do not make sense because it is not separate multiplicants of the terms of differential equations (9.9) and (9.10) which should be equated but the equations themselves.

To find $\Omega_i(\mathbf{r}_i)$ let us premultiply both equations (9.9) and (9.10) by Ψ_e^* and integrate over the co-ordinates of all electrons; then subtract (9.10) from (9.9). The right side turns zero. The first and the third terms also turn zero, and we obtain:

$$\frac{1}{2}\int \Psi_e^* \sum_{i,\, j\neq i} U_{ij}\Psi_e\, d\tau_e - \int \Psi_e^* \sum_i \Omega_i(\mathbf{r}_i)\,\Psi_e\, d\tau_e = 0 \qquad (9.12)$$

or

$$\sum_i \int \Psi_e^*\Omega_i(\mathbf{r}_i)\,\Psi_e\, d\tau_e = \sum_i \int \Psi_e^* \frac{1}{2} \sum_{j(j\neq i)} U_{ij}\Psi_e\, d\tau_e. \qquad (9.13)$$

Since the introduction of $\Omega_i(\mathbf{r}_i)$ reduces the electron problem to a problem for a system of non-interacting particles, Ψ_e may be represented by a product of wave functions of separate particles:

$$\Psi_e(\mathbf{r}_1, \mathbf{r}_2, \ldots) = \prod_i \psi_i(\mathbf{r}_i). \qquad (9.14)$$

Noting that $d\tau_e = d\tau_1\, d\tau_2 \ldots,$ we may rewrite (9.13) as follows:

$$\sum_i \int \psi_1^*(\mathbf{r}_1) \ldots \Omega_i(\mathbf{r}_i)\,\psi_1(\mathbf{r}_1) \ldots d\tau_1\, d\tau_2 \ldots =$$

$$= \sum_i \int \psi_i^*(\mathbf{r}_i)\,\Omega_i(\mathbf{r}_i)\,\psi_i(\mathbf{r}_i)\, d\tau_i =$$

$$= \frac{1}{2} \sum_{i,\, j\neq i} \int \psi_i^*(\mathbf{r}_1) \ldots U_{ij}(|\mathbf{r}_i - \mathbf{r}_j|)\,\psi_1(\mathbf{r}_1) \ldots d\tau_1\, d\tau_2 \ldots =$$

$$= \sum_i \int \psi_i^*(\mathbf{r}_i) \left[\frac{1}{2} \sum_{j\neq i} \int \psi_j^*(\mathbf{r}_j)\, U_{ij}(|\mathbf{r}_i - \mathbf{r}_j|)\, \psi_j(\mathbf{r}_j)\, d\tau_j \right] \psi_i\, d\tau_i. \qquad (9.15)$$

Comparing the second term of the chain of equations (9.15) with the last, we may write

$$\Omega_i(\mathbf{r}_i) = \frac{1}{2} \sum_{j\neq i} \int |\psi_j(\mathbf{r}_j)|^2 \frac{e^2}{4\pi\varepsilon_0 |\mathbf{r}_i - \mathbf{r}_j|}\, d\tau_j. \qquad (9.16)$$

The expression for $\Omega_i(\mathbf{r}_i)$ has the following meaning: $e|\psi_j(\mathbf{r}_j)|^2$ is the charge density of the electron cloud of the jth electron at the point $\mathbf{r}_j$; $e|\psi_j(\mathbf{r}_j)|^2 d\tau_j$ is the element of charge which determines the potential at point $\mathbf{r}_i$. Integrating over all the co-ordinates of the jth electron, we obtain the energy of interaction of the ith electron, with the "spread" jth electron.

After substitution of the expression (9.16) for $\Omega_i(\mathbf{r}_i)$ into (9.8) the equation for the functions $\psi_i(\mathbf{r}_i)$ becomes

$$-\frac{\hbar^2}{2m}\Delta_i\psi_i(\mathbf{r}_i)+\left[\frac{1}{2}\sum_{j\neq i}\int|\psi_j(\mathbf{r}_j)|^2\frac{e^2\,d\tau_j}{4\pi\varepsilon_0 r_{ij}}\right]\psi_i(\mathbf{r}_i)+$$
$$+U_i(\mathbf{r}_i,\ \mathbf{R}_1,\ \mathbf{R}_2,\ \ldots)\,\psi_i(\mathbf{r}_i)=E_i\psi_i(\mathbf{r}_i). \qquad (9.17)$$

To find $\Omega_i(\mathbf{r}_i)$ all the $\psi_j(\mathbf{r}_j)$ should be available, and the latter may be found only if all the $\Omega_i(\mathbf{r}_i)$ are known. *Equation* (9.17) *is called the Hartree equation*. It is an integro-differential equation. Its solution may be sought by the method of successive approximations. The procedure is as follows. Taking some functions $\psi_j^{(0)}(\mathbf{r}_j)$ as a zero approximation, calculate $\Omega_i^{(0)}(\mathbf{r}_i)$. Substitute $\Omega_i^{(0)}(\mathbf{r}_i)$ into (9.17) to obtain some functions $\psi_i^{(1)}(\mathbf{r}_i)$. Using them calculate $\Omega_i^{(1)}(\mathbf{r}_i)$, etc. *This procedure is repeated until the* $(n+1)$ *th approximation coincides with the nth within the limits of the predetermined error*. The shortcoming of the Hartree equation is that it takes no account of the Pauli exclusion principle. Introduction of the Pauli principle turns it into the Hartree-Fock equation. The Pauli principle requires the electron wave function to be antisymmetric under the interchange of any two electrons with the account of their co-ordinates and spin projections. At the same time $\prod_i \psi_i(\mathbf{r}_i)$ does not satisfy this condition. The right combination of wave functions of separate electrons takes the form of the Slater determinant

$$\Psi_e(\mathbf{q}_1,\ \mathbf{q}_2,\ \ldots)=\frac{1}{\sqrt{N!}}\begin{vmatrix}\psi_1(\mathbf{q}_1) & \psi_1(\mathbf{q}_2) & \ldots\\ \psi_2(\mathbf{q}_1) & \psi_2(\mathbf{q}_2) & \ldots\\ \ldots & \ldots & \ldots\end{vmatrix}, \qquad (9.18)$$

where N is the number of electrons; $\mathbf{q}_i$ is the set of four variables x_i, y_i, z_i and s_{zi}. The wave function Ψ_e satisfies the conditions:

$$\Psi_e(\ldots\mathbf{q}_i\ldots\mathbf{q}_k\ldots)=-\Psi_e(\ldots\mathbf{q}_k\ldots\mathbf{q}_i\ldots)$$
$$\int\Psi_e^*\Psi_e\,d\mathbf{q}_e=1. \qquad (9.19)$$

Using the expression for Ψ_e in the form of the Slater determinant, we obtain the expression for the energy E_i

$$E_i=\int\Psi_e^*(\mathbf{q}_1,\ \ldots)\left[-\frac{\hbar^2}{2m}\Delta_i+U_i(\mathbf{r}_i,\ \mathbf{R}_1,\ \mathbf{R}_2,\ \ldots)\right]\Psi_e d\mathbf{q}_e+$$
$$+\frac{1}{8\pi\varepsilon_0}\sum_{j\neq i}\int\Psi_e^*(\mathbf{q}_1,\ \ldots)\frac{e^2}{r_{ij}}\Psi_e(\mathbf{q}_1,\ \ldots)\,d\mathbf{q}_e, \qquad (9.20)$$

where $d\mathbf{q}_e$ is the volume element of the configurational space of the electron system which includes the spin variable. Integration over $d\mathbf{q}_e$ means both integration over the co-ordinates and summation over the spin variables of all the electrons.

The first integral in (9.20) coincides with the corresponding term of the Hartree equation, but the second term is different in that it contains exchange terms which are absent from the Hartree equation. This is because when integrating over $d\mathbf{q}_e$ we must retain terms containing the co-ordinates of the ith and jth electrons which now may be in any of the ψ_k and ψ_l and ψ_k' and ψ_l' states:

$$\frac{1}{8\pi\varepsilon_0}\sum_{j\neq i}\Psi_e^*(\mathbf{q}_1, \ldots)\frac{e^2}{r_{ij}}\Psi_e(\mathbf{q}_1, \ldots)\,d\mathbf{q}_e =$$

$$=\frac{1}{8\pi\varepsilon_0 N!}\sum_{j}\sum_{k,\,l}(-1)^{k+l}\int\psi_k^*(\mathbf{q}_i)\,\psi_l^*(\mathbf{q}_j)\times$$

$$\times\frac{e^2}{r_{ij}}\psi_k(\mathbf{q}_i)\,\psi_l(\mathbf{q}_j)\,d\mathbf{q}_i\,d\mathbf{q}_j. \qquad (9.21)$$

When $k=l$ we obtain the average Coulomb energy of electron interaction, when $k\neq l$, the exchange energy. The solution of the Schrödinger equation for a crystal by the Hartree-Fock method is, however, practically impossible.

Summary of Sec. 9

1. The motion of any given electron depends on the motion of all the other electrons, but since it itself affects the motion of other electrons, the motion of all electrons is self-consistent. This fact makes it possible to introduce the quantity Ω_i—the energy of the ith electron in the field of all the other electrons which takes into account the effect of the former on their motion. This field Ω_i is termed self-consistent.

2. The introduction of the self-consistent field reduces the equation for an electron system to a system of one-electron equations. The self-consistent field may, in principle, be obtained from the solution of the Hartree (9.17) or the Hartree-Fock (9.20) systems.

3. The introduction of the self-consistent field enables the electrons to be considered as non-interacting particles. Thus, the concept of the electron gas as an ideal gas is substantiated by quantum mechanics..

10. PERIODIC FIELD OF THE CRYSTAL LATTICE. TRANSLATIONAL OPERATOR

The introduction of the self-consistent field enables the problem of a system of interacting particles to be reduced to a one-particle problem. Using U without the index i for the potential energy of an arbitrary ith electron,

$$U = U(\mathbf{r}) = \Omega(\mathbf{r}) + U(\mathbf{r}, \mathbf{R}_1, \mathbf{R}_2, \ldots), \qquad (10.1)$$

we may represent the Hamiltonian of an arbitrary electron in the form

$$\hat{\mathrm{H}} = -\frac{\hbar^2}{2m}\Delta + U(\mathbf{r}). \qquad (10.2)$$

The energy E of the electron and its wave function may be found from the equation

$$\left[-\frac{\hbar^2}{2m}\Delta + U(\mathbf{r})\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}). \qquad (10.3)$$

It can be said of the potential energy $U(\mathbf{r})$ that in a crystal it should be a periodic function of the co-ordinates. Denote the lattice periods in three arbitrary directions by $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ and define the vector

$$\mathbf{n} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3, \qquad (10.4)$$

where n_1, n_2, n_3 are arbitrary integers. *The vector* $\mathbf{n}$ *is termed translation vector*. The periodicity condition assumes the form

$$U(\mathbf{r} + \mathbf{n}) = U(\mathbf{r}). \qquad (10.5)$$

Introduce the translational operator $\hat{\mathbf{T}}(\mathbf{n})$ which translates space by the vector $\mathbf{n}$ with the result that co-ordinates of every point $M(\mathbf{r})$ change by the value of $\mathbf{n}$. We therefore define $\hat{\mathbf{T}}(\mathbf{n})$ from the condition

$$\hat{\mathbf{T}}(\mathbf{n})\,\mathrm{f}(\mathbf{r}) = \mathrm{f}(\mathbf{r} + \mathbf{n}), \qquad (10.6)$$

where $\mathrm{f}(\mathbf{r})$ is an arbitrary function of the co-ordinates. Expanding $\mathrm{f}(\mathbf{r} + \mathbf{n})$ into the Taylor series at point $\mathbf{r}$, we obtain

$$\mathrm{f}(\mathbf{r} + \mathbf{n}) = \sum_{m=0}^{\infty}\frac{1}{m!}\frac{d^m\mathrm{f}(\mathbf{r})}{d\mathbf{r}^m}\mathbf{n}^m =$$

$$= \left(\sum_{m=0}^{\infty}\frac{1}{m!}\mathbf{n}^m\nabla^m\right)\mathrm{f}(\mathbf{r}) = e^{(\mathbf{n}\nabla)}\,\mathrm{f}(\mathbf{r}) = \hat{\mathbf{T}}(\mathbf{n})\,\mathrm{f}(\mathbf{r}) \left(\nabla \equiv \frac{d}{d\mathbf{r}}\right). \qquad (10.7)$$

The translational operator is therefore of the form

$$\hat{\mathbf{T}}(\mathbf{n}) = e^{(\mathbf{n}\nabla)}. \qquad (10.8)$$

The action of the translational operator on a periodic function produces the result

$$\hat{\mathbf{T}}(\mathbf{n})\,U(\mathbf{r}) = U(\mathbf{r} + \mathbf{n}) = U(\mathbf{r}). \qquad (10.9)$$

Hence, *an arbitrary periodic function is an eigenfunction of the operator* $\hat{\mathbf{T}}(\mathbf{n})$ *which corresponds to a unit eigenvalue.*

Should we discuss the action of the translational operator on the product of the potential energy (operator) and an arbitrary wave function, we would arrive at

$$\hat{\mathbf{T}}(\mathbf{n})\,\hat{\mathbf{U}}(\mathbf{r})\,\psi(\mathbf{r}) = \hat{\mathbf{T}}(\mathbf{n})\,U(\mathbf{r})\,\psi(\mathbf{r}) = U(\mathbf{r}+\mathbf{n})\,\psi(\mathbf{r}+\mathbf{n}) =$$

$$= U(\mathbf{r})\,\psi(\mathbf{r}+\mathbf{n}) = \hat{\mathbf{U}}(\mathbf{r})\,\hat{\mathbf{T}}(\mathbf{n})\,\psi(\mathbf{r}), \qquad (10.10)$$

$$\hat{\mathbf{T}}(\mathbf{n})\,\hat{\mathbf{U}} - \hat{\mathbf{U}}\hat{\mathbf{T}}(\mathbf{n}) = 0. \qquad (10.11)$$

In other words, *the translational operator commutes with the potential energy operator* $\hat{\mathbf{U}}(\mathbf{r}) \equiv U(\mathbf{r})$ *of the electron in a crystal.* It is easy to see that the kinetic energy operator, too, commutes with the translational operator:

$$\hat{\mathbf{T}}\hat{\mathbf{T}}(\mathbf{n}) = -\frac{\hbar^2}{2m}\nabla^2 \sum_{l=0}^{\infty} \frac{1}{l!}(\mathbf{n}^l\nabla^l) = \sum_{l=0}^{\infty} \frac{1}{l!}(\mathbf{n}^l\nabla^l)\left(-\frac{\hbar^2}{2m}\nabla^2\right) = \hat{\mathbf{T}}(\mathbf{n})\,\hat{\mathbf{T}}.$$

$$(10.12)$$

It follows from (10.11) and (10.12) that *the translational operator commutes with the Hamilton operator for a periodic potential field*

$$\hat{\mathbf{H}}\hat{\mathbf{T}}(\mathbf{n}) - \hat{\mathbf{T}}(\mathbf{n})\,\hat{\mathbf{H}} = 0, \qquad (10.13)$$

whence follows the conclusion that $\hat{\mathbf{T}}(\mathbf{n})$ *remains constant in time:*

$$\frac{d\hat{\mathbf{T}}(\mathbf{n})}{dt} = \frac{1}{i\hbar}\{\hat{\mathbf{T}}(\mathbf{n})\,\hat{\mathbf{H}} - \hat{\mathbf{H}}\hat{\mathbf{T}}(\mathbf{n})\} \equiv [\hat{\mathbf{H}},\ \hat{\mathbf{T}}(\mathbf{n})] = 0, \qquad (10.14)$$

and that *the system of wave functions of the operators* $\hat{\mathbf{T}}(\mathbf{n})$ *and* $\hat{\mathbf{H}}$ *is identical.*

Consider a reciprocal operator of $\hat{\mathbf{T}}(\mathbf{n})$ which we will denote by $\hat{\mathbf{T}}^{-1}(\mathbf{n})$. The product of the direct and the reciprocal operators should leave the space unchanged, or in other words the transformation should be the identity transformation as defined by the unit operator $\hat{\mathbf{1}}$:

$$\hat{\mathbf{T}}(\mathbf{n})\,\hat{\mathbf{T}}^{-1}(\mathbf{n}) \equiv \hat{\mathbf{T}}^{-1}(\mathbf{n})\,\mathbf{T}(\mathbf{n}) \equiv \hat{\mathbf{1}}, \qquad (10.15)$$

$$\hat{\mathbf{1}}\,\mathbf{f}(\mathbf{r}) \equiv \mathbf{f}(\mathbf{r}). \qquad (10.16)$$

It follows from the explicit expression for $\hat{\mathbf{T}}(\mathbf{n})$ that $\hat{\mathbf{T}}^{-1}(\mathbf{n})$ effects the translation of the space by the vector $-\mathbf{n}$:

$$\hat{\mathbf{T}}^{-1}(\mathbf{n}) = [e^{(\mathbf{n}\nabla)}]^{-1} = e^{(-\mathbf{n}\nabla)} = \hat{\mathbf{T}}(-\mathbf{n}). \qquad (10.17)$$

The square of the translational operator $\hat{T}^2(\mathbf{n})$ is the operator of translation by the vector $2\mathbf{n}$. By analogy

$$\hat{T}^s(\mathbf{n}) = \hat{T}(s\mathbf{n}). \qquad (10.18)$$

Suppose $\psi(\mathbf{r})$ is an eigenfunction of the translational operator corresponding to the eigenvalue $T(\mathbf{n})$:

$$\hat{T}(\mathbf{n})\psi(\mathbf{r}) = T(\mathbf{n})\psi(\mathbf{r}). \qquad (10.19)$$

Eigenvalues of $\hat{T}^2(\mathbf{n})$ are related to $T(\mathbf{n})$ by the condition

$$\hat{T}^2(\mathbf{n})\psi(\mathbf{r}) = \hat{T}(\mathbf{n})\,T(\mathbf{n})\psi(\mathbf{r}) = T^2(\mathbf{n})\psi(\mathbf{r}). \qquad (10.20)$$

It follows from (10.18) and (10.20) that

$$T^l(\mathbf{n}) = T(l\mathbf{n}), \qquad (10.21)$$

which is true of an exponential function.

Let us write $T(\mathbf{n})$ in the form

$$T(\mathbf{n}) = e^{i\varphi(\mathbf{n})}. \qquad (10.22)$$

Since translation by the vector n is a set of independent translations along the three axes, to fulfill the conditions (10.21) and (10.22) $\varphi(\mathbf{n})$ should be a linear scalar function:

$$\varphi(\mathbf{n}) = (\mathbf{kn}) = k_1 n_1 a_1 + k_2 n_2 a_2 + k_3 n_3 a_3. \qquad (10.23)$$

From the normalization condition

$$1 = \int |\psi(\mathbf{r}+\mathbf{n})|^2 d\tau = |e^{i\varphi(\mathbf{n})}|^2 \int |\psi(\mathbf{r})|^2 d\tau = |e^{i\varphi(\mathbf{n})}|^2 \qquad (10.24)$$

which is independent of the choice of the origin of co-ordinates it follows that $\varphi(\mathbf{n})$ is a real function, i. e., the *vector* **k** *should be real*. The dimensions of the **k** being inverse length, it became known as the wave vector. We will see later that dimensionality is not the only justification for such a name.

Thus,

$$\hat{T}(\mathbf{n})\psi(\mathbf{r}) = \psi(\mathbf{r}+\mathbf{n}) = e^{i(\mathbf{kn})}\psi(\mathbf{r}) = T(\mathbf{n})\psi(\mathbf{r}). \qquad (10.24')$$

As we know, the eigenfunctions of $\hat{T}(\mathbf{n})$ and $\hat{H}$ coincide, therefore we may write for the eigenfunctions of the Hamiltonian

$$\psi(\mathbf{r}+\mathbf{n}) = e^{i(\mathbf{kn})}\psi(\mathbf{r}). \qquad (10.25)$$

The condition (10.25) *is termed translational property of the wave function*. It may be said that eigenfunctions of the Hamiltonian for an electron moving in a periodic field satisfy the translational condition (10.25).

The vector $\mathbf{k}$ is characteristic of a specific wave function, therefore it should be shown as a subscript:

$$\psi(\mathbf{r}) = \psi_{\mathbf{k}}(\mathbf{r}). \qquad (10.26)$$

However, since

$$\hat{H}\psi_{\mathbf{k}}(\mathbf{r}) = E\psi_{\mathbf{k}}(\mathbf{r}), \qquad (10.27)$$

we may say that the *energy E should be a function of the wave vector, $E = E(\mathbf{k})$. The investigation of this dependence constitutes one of the major tasks of modern solid state physics.*

It follows from the above that the determination of the eigenfunctions of the translational operator involves the solution of the Schrödinger equation. In consequence their form should depend on the functional form of the potential energy $U(\mathbf{r})$. However, the general nature of the wave function may be assessed without solving the Schrödinger equation. With this end in view one should remember that the ∇ operator's eigenfunction has the form

$$\psi(\mathbf{r}) = Ae^{i(\mathbf{kr})} \qquad (10.28)$$

and satisfies the equation

$$\nabla\psi(\mathbf{r}) = i\mathbf{k}\psi(\mathbf{r}). \qquad (10.29)$$

At the same time it is the eigenfunction of the ∇^i operator and, consequently, of the e^{∇} operator. It is easily proved by direct substitution that the function $e^{i(\mathbf{kr})}$ is the eigenfunction of the translational operator:

$$\hat{T}(\mathbf{n})e^{i(\mathbf{kr})} = \left(\sum_{m=0}^{\infty}\frac{1}{m!}\mathbf{n}^m\nabla^m\right)e^{i(\mathbf{kr})} =$$

$$= \sum_{n=0}^{\infty}\frac{1}{m!}(\mathbf{n}\cdot i\mathbf{k})^m e^{i(\mathbf{kr})} = e^{i(\mathbf{kn})}e^{i(\mathbf{kr})} = T(\mathbf{n})e^{i(\mathbf{kr})}. \qquad (10.30)$$

However, this is not the only feasible form of the eigenfunction $\hat{T}(\mathbf{n})$ of the translational operator. An arbitrary periodical function $\varphi(\mathbf{r}+\mathbf{n}) = \varphi(\mathbf{r})$ multiplied by $e^{i(\mathbf{kr})}$ will, too, be an eigenfunction of $\hat{T}(\mathbf{n})$. The choice of the function $\varphi(\mathbf{r})$ is determined by the condition that $\psi_{\mathbf{k}}$ must also be the solution of the Schrödinger equation.

Thus, we arrive at an important conclusion that the *solution of the Schrödinger equation for the electron in a periodic field $U(\mathbf{r}+\mathbf{n}) = U(\mathbf{r})$ should be of the form*

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i(\mathbf{kr})}\varphi(\mathbf{r}) = e^{i(\mathbf{kr})}\varphi_{\mathbf{k}}(\mathbf{r}), \qquad (10.31)$$

where $\varphi_{\mathbf{k}}(\mathbf{r}+\mathbf{n})=\varphi_{\mathbf{k}}(\mathbf{r})$ *is a periodic function with the period of the potential field* $U(\mathbf{r})$; $e^{i(\mathbf{kr})}$ *is a plane wave propagating in the direction of the vector* $\mathbf{k}$. *The expression* (10.31) *for* $\psi_{\mathbf{k}}(\mathbf{r})$ *is called the Bloch wave* since the wave function $\psi_{\mathbf{k}}(\mathbf{r})$ may be visualized as a plane wave $e^{i(\mathbf{kr})}$ with a variable amplitude $\varphi_{\mathbf{k}}(\mathbf{r})$ modulated with the period of the crystal lattice. $\varphi(\mathbf{r})$ received the index $\mathbf{k}$ because $\varphi(\mathbf{r})$ may be different for different $\mathbf{k}$'s.

Summary of Sec. 10

1. The translational operator $\hat{\mathbf{T}}(\mathbf{n})$ translates the space by the translation vector $\mathbf{n}$ which results in the change of the space points co-ordinates in accordance with (10.6). $\hat{\mathbf{T}}(\mathbf{n})$ commutes with $\hat{\mathbf{H}}$ for a periodic field.

2. Eigenvalues of the translational operator are

$$T(\mathbf{n})=e^{i(\mathbf{kn})}, \tag{10.1s}$$

where $\mathbf{k}$ is an arbitrary real vector termed wave vector.

3. The solution of the Schrödinger equation for the electron in a periodic field has the form of the Bloch wave, or function:

$$\psi_{\mathbf{k}}(\mathbf{r})=e^{i(\mathbf{kr})}\varphi_{\mathbf{k}}(\mathbf{r}) \tag{10.2s}$$

subject to the condition

$$\varphi_{\mathbf{k}}(\mathbf{r}+\mathbf{n})=\varphi_{\mathbf{k}}(\mathbf{r}).$$

4. Electron energy is a function of the wave vector

$$E=E(\mathbf{k}). \tag{10.3s}$$

11. QUASIMOMENTUM

Many quantities considered in physics possess an important property—they obey conservation laws, i.e. in proper circumstances those quantities remain unchanged. Thus the momentum $\mathbf{p}=m\mathbf{v}$ of a particle moving through space with constant potential energy is conserved. The moment of momentum $\mathbf{M}=[\mathbf{rp}]$ is conserved in a centre-symmetrical field $U(\mathbf{r})=U(r)$. The energy of an isolated system is conserved if its Hamiltonian $\hat{\mathbf{H}}(\mathbf{p}, \mathbf{r})$ does not explicitly depend on time. These quantities are motion integrals in quantum mechanics as well. This fact is written in the operator representation in the following way: *a quantity L which does not explicitly depend on time is conserved if the corresponding operator* $\hat{\mathbf{L}}$ *commutes with the Hamilton operator*, since

$$\frac{d\hat{\mathbf{L}}}{dt}=[\hat{\mathbf{H}}, \hat{\mathbf{L}}]=\frac{1}{i\hbar}[\hat{\mathbf{L}}\hat{\mathbf{H}}-\hat{\mathbf{H}}\hat{\mathbf{L}}]. \tag{11.1}$$

The *physical basis of the conservation laws are certain properties of space and time symmetry.*

Actually, the momentum conservation law reflects the homogeneity of space; isotropy of space leads to the conservation of the moment of momentum, the homogeneity of time underlies the energy conservation law; the equivalence of the right and left screws results in the parity conservation law. In other words, a *definite symmetry of space and time results in the conservation of a certain physical quantity.* If some influence affects symmetry, the quantity corresponding to this symmetry will, too, be affected by this influence. Three of the above laws are easily deduced with the aid of the canonical Hamilton equations.

On the subject of electron motion in the periodical crystal lattice field we may make the following statement: *there must be some physical quantity that would be conserved during the motion of a particle in this field and that would correspond to the translational symmetry of the lattice potential field. Let us term this quantity quasimomentum.* The simplest reason for the term "quasimomentum" is its dimensions. Since translational symmetry reflects the invariability of the properties of space during its translation by any integral number of lattice periods, the dimensions of the quasimomentum should be the same as those of the momentum which reflects the homogeneity and the invariability of space in arbitrary displacements. It will be shown below that the properties of quasimomentum are to a large extent the same as those of the momentum. There should be an operator $\hat{\mathbf{P}}$ to correspond to the quasimomentum $\mathbf{P}$ that would commute with the lattice Hamiltonian:

$$\hat{\mathbf{P}}\hat{\mathbf{H}}-\hat{\mathbf{H}}\hat{\mathbf{P}}=0. \qquad (11.2)$$

We may thus say that eigenfunctions of the operators $\hat{\mathbf{H}}$ and $\hat{\mathbf{P}}$ for the case of the electron moving in the lattice field should coincide and that there should be a functional relation between their eigenvalues:

$$E=E\,(\mathbf{P}). \qquad (11.3)$$

In other words, *electron energy should be a function of the quasimomentum.*

It follows quite obviously from the commutation condition of $\hat{\mathbf{P}}$ and $\hat{\mathbf{H}}$ that the quasimomentum operator cannot be of the form $-i\hbar\nabla$ which is the form of the usual momentum operator, since the quasimomentum operator so defined does not commute with the Hamiltonian. This, by the way, is why the usual momentum of a particle moving in the lattice field is not conserved:

$$\frac{d\hat{\mathbf{p}}}{dt}=\frac{1}{i\hbar}\,(\hat{\mathbf{p}}\hat{\mathbf{H}}-\hat{\mathbf{H}}\hat{\mathbf{p}})=-(\nabla U)=\mathbf{F}_{l}. \qquad (11.4)$$

At the same time it is obvious that there should be some connection between $\hat{\mathbf{p}}$ and $\hat{\mathbf{P}}$. To prove it suppose $\nabla U \rightarrow 0$, i.e. the potential energy of the lattice field tends to a constant. In this limiting case quasimomentum and momentum should be identical. This means that the quasimomentum operator should contain some variable which depends on the form of the potential field $U(\mathbf{r})$ and tends to zero as $\nabla U \rightarrow 0$. This enables us to write for $\hat{\mathbf{P}}$:

$$\hat{\mathbf{P}} = -i\hbar\nabla + i\hbar\hat{\gamma}(\mathbf{r}), \tag{11.5}$$

where $\hat{\gamma}(\mathbf{r}) \rightarrow 0$ as $\nabla U \rightarrow 0$. $\hat{\gamma}(\mathbf{r})$ should provide for the commutation of $\hat{\mathbf{P}}$ and $\hat{\mathbf{H}}$.

We will try to find the form of the operator $\hat{\mathbf{P}}$ from the eigenfunctions and eigenvalues equation taking into account that $\psi_{\mathbf{k}}$ is an eigenfunction of the quasimomentum operator:

$$\hat{\mathbf{P}}\psi_{\mathbf{k}}(\mathbf{r}) = \mathbf{P}\psi_{\mathbf{k}}(\mathbf{r}). \tag{11.6}$$

To find from here the operator $\hat{\mathbf{P}}$ we will have to represent $\psi_{\mathbf{k}}$ in the form of a Bloch wave and $\hat{\mathbf{P}}$—in the form of $-i\hbar\nabla + i\hbar\gamma$. We will have an equation for $\hat{\gamma}$:

$$\begin{aligned}\hat{\mathbf{P}}\psi_{\mathbf{k}}(\mathbf{r}) &= -i\hbar i\mathbf{k}\psi_{\mathbf{k}}(\mathbf{r}) + e^{i(\mathbf{kr})}(-i\hbar\nabla\varphi_{\mathbf{k}}(\mathbf{r})) + i\hbar\hat{\gamma}\psi_{\mathbf{k}}(\mathbf{r}) = \\ &= \hbar\mathbf{k}\psi_{\mathbf{k}}(\mathbf{r}) + i\hbar[\hat{\gamma} - \nabla\ln\varphi_{\mathbf{k}}(\mathbf{r})]\psi_{\mathbf{k}}(\mathbf{r}) = \mathbf{P}\psi_{\mathbf{k}}(\mathbf{r}).\end{aligned} \tag{11.7}$$

From (11.7) it follows that

$$\mathbf{P} = \hbar\mathbf{k} \tag{11.8}$$

and

$$\gamma = (\nabla\ln\varphi_{\mathbf{k}}(\mathbf{r})). \tag{11.9}$$

We see from here that $\hat{\gamma}$ is a multiplication operator and that it depends on the form of the potential field $U(\mathbf{r})$ through the dependence on the periodic function $\varphi_{\mathbf{k}}(\mathbf{r})$. As $\nabla U(\mathbf{r})$ tends to zero $\varphi_{\mathbf{k}}(\mathbf{r})$ tends to a constant and $\hat{\gamma}$ tends to zero. This results in the quasimomentum being in this limiting case identical with the normal momentum $\hat{\mathbf{p}}$.

Hence, the *quasimomentum operator is of the form*

$$\hat{\mathbf{P}}(\mathbf{r}) = -i\hbar\nabla + i\hbar\,(\vec{\nabla}\ln\varphi_{\mathbf{k}}(\mathbf{r})). \tag{11.10}$$

Suppose now that some additional field $V(\mathbf{r})$ with a different periodicity is superimposed on the periodic field $U(\mathbf{r})$. In this case

$$\hat{\mathbf{H}} = \hat{\mathbf{T}} + \hat{\mathbf{U}} + \hat{\mathbf{V}} = \hat{\mathbf{H}}_0 + \hat{\mathbf{V}}. \tag{11.11}$$

Since $\hat{\mathbf{P}}$ commutes with the lattice field Hamiltonian $\hat{\mathbf{H}}_0$ and $(\nabla \ln \varphi_{\mathbf{k}})$ commutes with $\hat{V}$ (for both are multiplication operators), the time derivative of quasimomentum will be equal to the external force $\hat{\mathbf{F}}_a$:

$$\hat{\mathbf{F}}_a = -(\nabla V). \qquad (11.12)$$

To prove it write

$$\frac{d\hat{\mathbf{P}}}{dt} = [\hat{\mathbf{H}}\hat{\mathbf{P}}] = [\hat{\mathbf{V}}\hat{\mathbf{P}}] = \frac{1}{i\hbar}\{(-i\hbar\nabla)V - V(-i\hbar\nabla)\} = -(\nabla V) = \mathbf{F}_a. \qquad (11.13)$$

Thus, the *non-periodic part of the potential field*—$|\nabla V(\mathbf{r})|$ *changes the quasimomentum*. This means that *any deviations frcm the ideal lattice field cause changes in the quasimomentum* $\mathbf{P}$ and, consequently, *any defects in the ideal lattice occasion the scattering of electron waves*. Such deviations of $U(\mathbf{r})$ from periodicity are caused by thermal vibrations and lattice imperfections. *Electron scattering on these defects is the physical reason for electric conductivity being finite*. If an external force field $V(\mathbf{r})$ is applied to the ideal crystal, the quasimomentum will change only under the influence of external force $\mathbf{F}_a$; the momentum, on the other hand, changes under the combined influence of external $\mathbf{F}_a$ and internal $\mathbf{F}_i = -\nabla U(\mathbf{r})$ forces:

$$\frac{d\mathbf{p}}{dt} = \mathbf{F}_i + \mathbf{F}_a. \qquad (11.14)$$

The energy of the electron should depend on the wave vector $\mathbf{k}$ or the quasimomentum $\mathbf{P} = \hbar\mathbf{k}$. The specific form of the functional dependence $E(\mathbf{k})$ or $E(\mathbf{P})$ can only be found after the Schrödinger equation

$$\hat{\mathbf{H}}\psi_{\mathbf{k}}(\mathbf{r}) = E(\mathbf{k})\,\psi_{\mathbf{k}}(\mathbf{r}) \qquad (11.15)$$

is solved.

Let us find the equation which $\varphi_{\mathbf{k}}$ should satisfy. To this end we take into account that

$$\nabla\psi_{\mathbf{k}}(\mathbf{r}) = i\mathbf{k}\varphi_{\mathbf{k}}(\mathbf{r}) + e^{i(\mathbf{kr})}[\nabla\varphi_{\mathbf{k}}(\mathbf{r})]; \qquad (11.16)$$

or

$$\Delta\psi_{\mathbf{k}}(\mathbf{r}) = i\mathbf{k}\,[i\mathbf{k}\psi_{\mathbf{k}}(\mathbf{r}) + e^{i(\mathbf{kr})}(\nabla\varphi_{\mathbf{k}}(\mathbf{r}))] + \\ + i\mathbf{k}e^{i(\mathbf{kr})}[\nabla\varphi_{\mathbf{k}}(\mathbf{r})] + e^{i(\mathbf{kr})}[\Delta\varphi_{\mathbf{k}}(\mathbf{r})]. \qquad (11.17)$$

Hence, substituting (11.17) into the equation (11.15) and cancelling out $e^{i(\mathbf{kr})}$, we obtain an equation for $\varphi_{\mathbf{k}}(\mathbf{r})$:

$$-\frac{\hbar^2}{2m}[-\mathbf{k}^2\varphi_{\mathbf{k}}(\mathbf{r}) + 2i(\mathbf{k}\cdot\nabla\varphi_{\mathbf{k}}(\mathbf{r})) + \Delta\varphi_{\mathbf{k}}(\mathbf{r})] + U(\mathbf{r})\varphi_{\mathbf{k}} = E(\mathbf{k})\,\varphi_{\mathbf{k}}(\mathbf{r}). \qquad (11.18)$$

or

$$-\frac{\hbar^2}{2m}\Delta\varphi_{\mathbf{k}}(\mathbf{r})+\left[\frac{\hbar^2\mathbf{k}^2}{2m}-\frac{i\hbar^2}{m}(\mathbf{k}\nabla)+U(\mathbf{r})\right]\varphi_{\mathbf{k}}(\mathbf{r})=E(\mathbf{k})\,\varphi_{\mathbf{k}}(\mathbf{r}). \quad (11.19)$$

The equation (11.19) shows that $\varphi_{\mathbf{k}}(\mathbf{r})$ depends on the value of $\mathbf{k}$; hence, $\mathbf{k}$ is used as a subscript. Since energy is a real function, i.e.

$$E^*(\mathbf{k})=E(\mathbf{k}), \quad (11.20)$$

the Schrödinger equation for the complex conjugate wave function $\psi_{\mathbf{k}}^*(\mathbf{r})$ may be written in the form

$$\hat{\mathbf{H}}^*\psi_{\mathbf{k}}^*(\mathbf{r})=E(\mathbf{k})\,\psi_{\mathbf{k}}^*(\mathbf{r}). \quad (11.21)$$

Taking into account that

$$\hat{\mathbf{H}}^*=-\frac{\hbar^2}{2m}\Delta+U(\mathbf{r}) \quad (11.22)$$

and

$$\psi_{\mathbf{k}}^*(\mathbf{r})=e^{-i(\mathbf{kr})}\varphi_{\mathbf{k}}^*(\mathbf{r}), \quad (11.23)$$

we obtain for $\varphi_{\mathbf{k}}^*(\mathbf{r})$

$$-\frac{\hbar^2}{2m}\Delta\varphi_{\mathbf{k}}^*(\mathbf{r})+\left[\frac{\hbar^2\mathbf{k}^2}{2m}+\frac{i\hbar^2}{m}(\mathbf{k}\nabla)+U(\mathbf{r})\right]\varphi_{\mathbf{k}}^*(\mathbf{r})=E(\mathbf{k})\,\varphi_{\mathbf{k}}^*(\mathbf{r}). \quad (11.24)$$

Let us now write the equation (11.19) for the function with the vector $(-\mathbf{k})$

$$-\frac{\hbar^2}{2m}\Delta\varphi_{-\mathbf{k}}(\mathbf{r})+\left[\frac{\hbar^2\mathbf{k}^2}{2m}+\frac{i\hbar^2}{m}(\mathbf{k}\nabla)+U(\mathbf{r})\right]\varphi_{-\mathbf{k}}(\mathbf{r})=E(-\mathbf{k})\,\varphi_{-\mathbf{k}}(\mathbf{r}). \quad (11.25)$$

If $\varphi_{\mathbf{k}}^*(\mathbf{r})=\varphi_{-\mathbf{k}}(\mathbf{r})$, this means that equations (11.24) and (11.25) coincide, and consequently the condition

$$E(\mathbf{k})=E(-\mathbf{k}) \quad (11.26)$$

is fulfilled, i. e. *energy is an even function of the wave vector*. In the vicinity of $\mathbf{k}=0$ energy depends at least on $\mathbf{k}^2$.

In the $k_xk_yk_z$-space the *equation*

$$E(\mathbf{k})=\text{const} \quad (11.27)$$

represents some surface—a surface of constant energy. The form of constant-energy surfaces determines many properties of semiconductors.

Summary of Sec. 11

1. There should be a conservative physical quantity to correspond to the translational symmetry of the lattice field. This quantity is termed quasimomentum.

2. .The quasimomentum operator commutes with the lattice field Hamiltonian. Its eigenfunctions are the Bloch functions $\psi_{\mathbf{k}}(\mathbf{r})$. The eigenvalues $\mathbf{P}$ are related to the $\mathbf{k}$ vector

$$\mathbf{P}=\hbar\mathbf{k} \tag{11.1s}$$

and the operator $\hat{\mathbf{P}}$ itself is of the form

$$\hat{\mathbf{P}}=-i\hbar\nabla+i\hbar[\nabla\ln\varphi_{\mathbf{k}}(\mathbf{r})]. \tag{11.2s}$$

3. Energy is a function of the quasimomentum and of the wave vector

$$E=E(\mathbf{P});\quad E=E(\mathbf{k}) \tag{11.3s}$$

and the equation

$$E(\mathbf{P})=\text{const, or } E(\mathbf{k})=\text{const}, \tag{11.4s}$$

defines a surface in the $\mathbf{P}$, or $\mathbf{k}$, space which is termed constant-energy surface.

4. If a field $V(\mathbf{r})$ with a periodicity different from that of the lattice is applied to the crystal, the quasimomentum will change in accordance with the equation

$$\frac{d\mathbf{P}}{dt}=-\nabla V(\mathbf{r})=\mathbf{F}_a. \tag{11.5s}$$

At the same time the momentum $\mathbf{p}$ changes under the influence both of the external forces $\mathbf{F}_a$ and the forces of the lattice field $\mathbf{F}_l=-\nabla U(\mathbf{r})$:

$$\frac{d\mathbf{p}}{dt}=\mathbf{F}_a+\mathbf{F}_l. \tag{11.6s}$$

12. THE EFFECTIVE MASS OF THE ELECTRON

Suppose $\mathbf{k}_0$ (or $\mathbf{P}_0$) is an extremal point

$$E(\mathbf{k}_0)=E_0=E_{extr}. \tag{12.1}$$

In addition to $\mathbf{k}_0$, there should be other extremal points, for instance, the point $-\mathbf{k}_0$ symmetrical to it. Since energy exhibits parity not only with respect to $\mathbf{k}$ but its projections as well

$$E(k_x, k_y, k_z)=E(-k_x, k_y, k_z)=\ldots, \tag{12.2}$$

it may be said that the *number of extremal points should be determined by the symmetry elements of the lattice field, i.e., of potential energy*. This can be proved by subjecting the Hamiltonian to a symmetry transformation. In a cubic lattice with its 24 symmetry elements the number of equivalent extremal points can in general be 24.

Let us expand $E(\mathbf{k})$ into the Taylor series around one of the extremal points $\mathbf{k}_0$:

$$E(\mathbf{k})=\sum_{l=0}^{\infty}\frac{d^lE(\mathbf{k})}{a\mathbf{k}^l}\Bigg|_{\mathbf{k}_0}\frac{(\mathbf{k}-\mathbf{k}_0)^l}{l!}=E(\mathbf{k}_0)+\frac{dE(\mathbf{k}_0)}{d\mathbf{k}}(\mathbf{k}-\mathbf{k}_0)+\dots . \quad (12.3)$$

Differentiating with respect to the vector argument, $\frac{d}{d\mathbf{k}}$, gives a set of three quantities obtained by differentiating with respect to k_x, k_y, k_z; therefore *the symbolic derivative in the product is in reality a complex expression termed covariant tensor of rank l.* When $l=0$ we obtain a scalar, when $l=1$, a tensor of rank one which is a vector; when $l=2$, 3, etc., a tensor of two, three, etc. Let us write out only the first two terms for $l=1$ and $l=2$:

$$\frac{dE}{d\mathbf{k}}=\left(\frac{\partial E}{\partial k_x},\ \frac{\partial E}{\partial k_y},\ \frac{\partial E}{\partial k_z}\right); \quad (12.4)$$

$$\frac{d^2E}{d\mathbf{k}^2}=\left(\frac{d}{d\mathbf{k}}\frac{\partial E}{\partial k_x},\ \frac{d}{d\mathbf{k}}\frac{\partial E}{\partial k_y},\ \frac{d}{d\mathbf{k}}\frac{\partial E}{\partial k_z}\right)=$$

$$=\begin{bmatrix} \frac{\partial^2E}{\partial k_x^2}; & \frac{\partial^2E}{\partial k_y\,\partial k_x}; & \frac{\partial^2E}{\partial k_z\,\partial k_x} \\ \frac{\partial^2E}{\partial k_x\,\partial k_y}; & \frac{\partial^2E}{\partial k_y^2}; & \frac{\partial^2E}{\partial k_z\,\partial k_y} \\ \frac{\partial^2E}{\partial k_x\,\partial k_z}; & \frac{\partial^2E}{\partial k_y\,\partial k_z}; & \frac{\partial^2E}{\partial k_z^2} \end{bmatrix}. \quad (12.5)$$

Thus, $\frac{d^2E}{d\mathbf{k}^2}$ is made up of nine second-order partial derivatives with respect to the wave vector. The quantities $\frac{\partial^2E}{\partial k_i\,\partial k_j}$ are termed components, or elements, of the tensor. The value of a mixed derivative is independent of the order of differentiation,

$$\frac{\partial^2E}{\partial k_i\,\partial k_j}=\frac{\partial^2E}{\partial k_j\,\partial k_i}, \quad (12.6)$$

and this makes the tensor symmetrical; $\frac{\partial^2E}{\partial k_i^2}$ are its diagonal elements. The derivative of order l constitutes a tensor of rank l with 3^l elements.

Consider a small element of space around the point $\mathbf{k}_0$. In this case only first several terms of the series may be considered. Since the expansion into the Taylor series was performed around the extremal point, $\frac{dE}{d\mathbf{k}}\Big|_{\mathbf{k}_0}=0$ (the condition for the extremum), and the series begins with quadratic terms:

$$E(\mathbf{k}) = E_0 + \frac{1}{2}\frac{d^2E}{d\mathbf{k}^2}(\mathbf{k}-\mathbf{k}_0)^2 + \ldots = E_0 + \\ + \frac{1}{2}\sum_{i,j}\frac{\partial^2 E}{\partial k_i\,\partial k_j}(k_i - k_{0i})(k_j - k_{0j}) + \ldots . \tag{12.7}$$

We see from here that the *constant-energy surface near an extremal point is sufficiently accurately represented by a second-order surface.* The accuracy will be the greater the nearer the energy value to the extreme value E_0. This will be the case when the neglected terms will be small as compared to the first; for instance,

$$\frac{1}{6}\frac{d^3E}{\partial\mathbf{k}^3}(\mathbf{k}-\mathbf{k}_0)^3 \ll \frac{1}{2}\frac{d^2E}{\partial\mathbf{k}^2}(\mathbf{k}-\mathbf{k}_0)^2. \tag{12.8}$$

A tensor of rank two may be, by an appropriate choice of co-ordinate axes, transformed into a diagonal tensor, i.e. the non diagonal elements for the new axes will turn zero. Suppose we found such axes that

$$\frac{\partial^2 E}{\partial k_i\,\partial k_j} = 0 \text{ for } i \neq j. \tag{12.9}$$

Now the equation for the constant-energy surface assumes the form

$$E(\mathbf{k}) = E(\mathbf{k}_0) + \frac{1}{2}\sum_{i=1}^{3}\frac{\partial^2 E}{\partial k_i^2}\cdot(k_i - k_{0i})^2 = \text{const} = E \tag{12.10}$$

or

$$E(\mathbf{k}) - E_0 = \frac{1}{2}\frac{\partial^2 E}{\partial k_x^2}(k_x - k_{0x})^2 + \frac{1}{2}\frac{\partial^2 E}{\partial k_y^2}(k_y - k_{0y})^2 + \\ + \frac{1}{2}\frac{\partial^2 E}{\partial k_z^2}(k_z - k_{0z})^2 = \text{const}'. \tag{12.11}$$

Since the expansion is performed around the extremal point the sign of all the three derivatives is the same: plus in a minimum and minus in a maximum; therefore, *the constant-energy surface is an ellipsoid.*

Consider the form of constant-energy surfaces in the quasimomentum space. Obviously,

$$E(\mathbf{P}) = E(\mathbf{P}_0) + \frac{1}{2}\frac{d^2E}{d\mathbf{P}^2}\cdot(\mathbf{P}-\mathbf{P}_0)^2 + \frac{1}{6}\frac{d^3E}{d\mathbf{P}^3}\cdot(\mathbf{P}-\mathbf{P}_0)^3 + \ldots, \tag{12.12}$$

since $\mathbf{P}_0 = \hbar\mathbf{k}_0$ is an extremal point. In a sufficiently small vicinity of the point $\mathbf{P}_0$ we may write

$$E(\mathbf{P}) = E_0 + \frac{1}{2}\frac{d^2E}{d\mathbf{P}^2}\cdot(\mathbf{P}-\mathbf{P}_0)^2 + \ldots . \tag{12.13}$$

Set

$$\frac{d^2E}{d\mathbf{P}^2}=\mathbf{m}^{*-1}. \qquad (12.14)$$

Evidently, the elements of the tensor $\mathbf{m}^{*-1}$ are

$$m_{ij}^{*-1}=\frac{\partial^2 E}{\partial P_i \partial P_j}. \qquad (12.15)$$

Since the dimensions of quasimomentum and momentum are the same, the dimensions of the components of tensor $\mathbf{m}^{*-1}$ are those of reciprocal mass; in other words, the dimensions of $[m_{ij}^*]=\left[\frac{\partial^2 E}{\partial P_i \partial P_j}\right]^{-1}=$ $=[M]$ are the dimensions of mass. *The quantity* $\mathbf{m}^{*-1}=\frac{d^2E}{d\mathbf{P}^2}$ *is termed the reciprocal effective mass tensor*. The expression for energy in terms of the reciprocal effective mass tensor may be written in the form:

$$E(\mathbf{P})=E_0+\frac{1}{2\mathbf{m}^*}\cdot(\mathbf{P}-\mathbf{P}_0)^2, \qquad (12.16)$$

which looks like the expression for kinetic energy of a free particle having the momentum $\mathbf{p}-\mathbf{p}_0$.

Note that *the effective mass is positive in the minimum and negative in the maximum of the energy*. The general form of the expression for $\mathbf{m}^{*-1}$ is:

$$\mathbf{m}^{*-1}=\begin{pmatrix} m_{xx}^{-1} & m_{xy}^{-1} & m_{xz}^{-1} \\ m_{yx}^{-1} & m_{yy}^{-1} & m_{yz}^{-1} \\ m_{zx}^{-1} & m_{zy}^{-1} & m_{zz}^{-1} \end{pmatrix}=\begin{bmatrix} \frac{\partial^2 E}{\partial P_x^2} & \frac{\partial^2 E}{\partial P_x \partial P_y} & \frac{\partial^2 E}{\partial P_x \partial P_z} \\ \frac{\partial^2 E}{\partial P_y \partial P_x} & \frac{\partial^2 E}{\partial P_y^2} & \frac{\partial^2 E}{\partial P_y \partial P_z} \\ \frac{\partial^2 E}{\partial P_z \partial P_x} & \frac{\partial^2 E}{\partial P_z \partial P_y} & \frac{\partial^2 E}{\partial P_z^2} \end{bmatrix}. \qquad (12.17)$$

When the tensor is diagonalized,

$$\mathbf{m}^{*-1}=\frac{1}{\mathbf{m}^*}=\begin{pmatrix} m_{xx}^{-1} & 0 & 0 \\ 0 & m_{yy}^{-1} & 0 \\ 0 & 0 & m_{zz}^{-1} \end{pmatrix}=\begin{pmatrix} m_1^{-1} & 0 & 0 \\ 0 & m_2^{-1} & 0 \\ 0 & 0 & m_3^{-1} \end{pmatrix}, \qquad (12.18)$$

where

$$m_i^{-1}=\frac{\partial^2 E}{\partial P_i^2}=m_{ii}^{-1}. \qquad (12.19)$$

Introduce a tensor reciprocal of the reciprocal effective mass tensor

$$\{\mathbf{m}^{*-1}\}^{-1}=\mathbf{m}^*. \qquad (12.20)$$

On the grounds of its dimensions, we will term it *the effective mass tensor*. Its components m_{ij} are not equal to the reciprocal components of the reciprocal effective mass tensor, i.e.

$$m_{ij} \neq \frac{1}{\frac{\partial^2 E}{\partial P_i \partial P_j}}. \tag{12.21}$$

They should be found from the condition

$$\mathbf{m}^* \mathbf{m}^{*-1} = \mathbf{m}^{*-1} \mathbf{m}^* = \mathbf{I}, \tag{12.22}$$

where $\mathbf{I}$ is the unit tensor. Consider a specific case when tensor $\mathbf{m}^{*-1}$ is diagonal:

$$\{\mathbf{m}^{*-1}\}_{ij} = m_i^{-1}\delta_{ij}. \tag{12.23}$$

Denote the components of the tensor $\mathbf{m}^*$ by m_{ij}. Since

$$I_{ij} = \delta_{ij} = \begin{cases} 1 \text{ for } i=j \\ 0 \text{ for } i \neq j, \end{cases} \tag{12.24}$$

we rewrite (12.22), with due regard to (12.24) and by the rules of matrix multiplication,

$$\sum_l \mathbf{m}_{il} \{\mathbf{m}^{*-1}\}_{lj} = \sum_l m_{il} m_l^{-1} \delta_{lj} = m_{ij} m_j^{-1} = \delta_{ij}. \tag{12.25}$$

We see that

$$m_{ij} = \frac{1}{m_i^{-1}} \delta_{ij} = m_j \delta_{ij}, \tag{12.26}$$

i.e. *if tensor* $\mathbf{m}^{*-1}$ *is a diagonal tensor, then tensor* $\mathbf{m}^*$ *will be diagonal, too, and its diagonal components will be reciprocal of the components of the reciprocal effective mass tensor*. We will write it out in detail. Denoting

$$m_{ii}^{-1} = \frac{\partial^2 E}{\partial P_i^2} = m_i^{-1}, \tag{12.27}$$

we obtain the tensors $\mathbf{m}^*$ and $\mathbf{m}^{*-1}$ in the form

$$\mathbf{m}^{*-1} = \begin{bmatrix} \frac{1}{m_1} & 0 & 0 \\ 0 & \frac{1}{m_2} & 0 \\ 0 & 0 & \frac{1}{m_3} \end{bmatrix}; \qquad \mathbf{m}^* = \begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{pmatrix}. \tag{12.28}$$

The quantities m_i will be termed below effective mass components. Division by $\mathbf{m}^*$ will be understood to mean multiplication by $\mathbf{m}^{*-1}$ (and vice versa).

The equation for the constant-energy surface is

$$E(\mathbf{P})=E_0+\frac{(P_x-P_{0x})^2}{2m_1}+\frac{(P_y-P_{0y})^2}{2m_2}+\frac{(P_z-P_{0z})^2}{2m_3}=\text{const}. \quad (12.29)$$

Introducing the semi-axes of the ellipsoid a, b, c and writing out the ellipsoid equation in the canonical form

$$\frac{(P_x-P_{0x})^2}{a^2}+\frac{(P_y-P_{0y})^2}{b^2}+\frac{(P_z-P_{0z})^2}{c^2}=1, \quad (12.30)$$

we obtain the obvious result that the *length of a semi-axis is proportional to the square root of the corresponding m_i:*

$$a^2=2(E-E_0)m_1;\ b^2=2(E-E_0)m_2;\ c^2=2(E-E_0)m_3. \quad (12.31)$$

Consider a case when all the three effective mass components are equal: $m_1=m_2=m_3=m^*$.

In this case the *effective mass tensor degenerates into a scalar* (a tensor of rank zero), *and constant-energy surfaces take the form of spheres:*

$$E=E_0+\frac{(\mathbf{P}-\mathbf{P}_0)^2}{2m^*}=\text{const}. \quad (12.32)$$

If any two components of the tensor are equal, for example, $m_1=m_2\neq m_3$, then

$$E=E_0+\frac{(P_x-P_{0x})^2+(P_y-P_{0y})^2}{2m_1}+\frac{(P_z-P_{0z})^2}{2m_3}=\text{const}, \quad (12.33)$$

and *constant-energy surfaces take the form of ellipsoids of revolution with the rotation axis along P_z.* If $m_1<m_3$, the ellipsoid of revolution will be elongated along the rotation axis; the elongation being the greater, the greater is the ratio of the effective masses $\frac{m_3}{m_1}$. If $m_1>m_3$, the ellipsoid will be contracted along the rotation axis.

In the general case of different effective masses m_i the ellipsoid will have three different semi-axes. It must be remembered that the greater is the difference in the effective masses m_1, m_2, m_3, the more will the ellipsoid be "elongated".

As the energy difference $E(\mathbf{k})-E_0$. increases, further terms of the Taylor series must be taken into account. The energy ellipsoids will be deformed, being transformed into more complex energy surfaces.

13. RELATION BETWEEN VELOCITY AND QUASIMOMENTUM

The velocity operator $\hat{\mathbf{v}}$ is defined with the aid of quantum Poisson brackets

$$\hat{\mathbf{v}} = \frac{d\hat{\mathbf{r}}}{dt} = [\hat{\mathbf{H}}, \hat{\mathbf{r}}] = \frac{1}{i\hbar}(\hat{\mathbf{r}}\hat{\mathbf{H}} - \hat{\mathbf{H}}\hat{\mathbf{r}}), \tag{13.1}$$

where $\hat{\mathbf{r}}$ is the co-ordinate operator, and $\hat{\mathbf{H}}$ the Hamilton operator. To calculate the commutator $\hat{\mathbf{r}}$ and $\hat{\mathbf{H}}$ it is convenient to go over to the E- or **k**-representation of specific operations performed with the variable **k** on which all the functions may depend. *The Hamilton operator in* **k**- *or E-representation is a multiplication operator*, or simply the energy $E(\mathbf{k})$:

$$\hat{\mathbf{H}}(\mathbf{k}) \equiv E(\mathbf{k}). \tag{13.2}$$

To deduce the form of the operator $\hat{\mathbf{r}}(\mathbf{k})$ we will have to take into account that $\psi_{\mathbf{k}}(\mathbf{r})$ is closely related to the eigenfunction of the operator $\hat{\mathbf{r}}(\mathbf{k})$, which we denote by $\psi_{\mathbf{r}}(\mathbf{k})$:

$$\hat{\mathbf{r}}(\mathbf{k})\,\psi_{\mathbf{r}}(\mathbf{k}) = \mathbf{r}\psi_{\mathbf{r}}(\mathbf{k}), \tag{13.3}$$

where **r** is the eigenvalue of the co-ordinate operator, and $\psi_{\mathbf{r}}(\mathbf{k})$, its eigenfunction in the **k**-representation.

It is established in quantum mechanics that eigenfunctions of two operators in mutual representations are related by a simple relation: if the eigenfunction of $\hat{\mathbf{L}}$ in M-representation is $\psi_L(M)$,

$$\hat{\mathbf{L}}(M)\,\psi_L(M) = L\psi_L(M), \tag{13.4}$$

and the eigenfunction of $\hat{\mathbf{M}}$ in L-representation is $\psi_M(L)$,

$$\hat{\mathbf{M}}(L)\,\psi_M(L) = M\psi_M(L), \tag{13.5}$$

then

$$\psi_L(M) = \psi_M^*(L). \tag{13.6}$$

This enables us to write *the eigenfunction of the operator* $\hat{\mathbf{r}}(\mathbf{k})$ *in* **k**-*representation in the form*

$$\psi_{\mathbf{r}}(\mathbf{k}) = \psi_{\mathbf{k}}^*(\mathbf{r}) = e^{-i(\mathbf{kr})}\,\varphi_{\mathbf{k}}^*(\mathbf{r}). \tag{13.7}$$

After that it will be easy to find the form of the $\hat{\mathbf{r}}(\mathbf{k})$ operator with the aid of the eigenfunctions and eigenvalues equation:

$$\hat{\mathbf{r}}(\mathbf{k})\,\psi_{\mathbf{r}}(\mathbf{k}) = \mathbf{r}\psi_{\mathbf{r}}(\mathbf{k}), \tag{13.8}$$

or

$$\hat{\mathbf{r}}(\mathbf{k})\,\psi_{\mathbf{k}}^*(\mathbf{r}) = \mathbf{r}\psi_{\mathbf{k}}^*(\mathbf{r}). \tag{13.9}$$

This relation means that we must find such a form of the $\hat{\mathbf{r}}(\mathbf{k})$ operator which would enable us to obtain as a result of its action on the function $\psi_{\mathbf{k}}^{*}(\mathbf{r})$ the original function multiplied by $\mathbf{r}$. The action of the operator on the function $\psi_{\mathbf{k}}^{*}(\mathbf{r})$ is to be limited to the variable $\mathbf{k}$.

Consider the action of the operator $\frac{d}{d\mathbf{k}} = \nabla_{\mathbf{k}}$ on the function $\psi_{\mathbf{k}}^{*}(\mathbf{r})$:

$$\nabla_{\mathbf{k}}\psi_{\mathbf{k}}^{*}(\mathbf{r}) = \nabla_{\mathbf{k}}[e^{-i(\mathbf{kr})}\varphi_{\mathbf{k}}^{*}(\mathbf{r})] = -i\mathbf{r}\psi_{\mathbf{k}}^{*}(\mathbf{r}) + \\ + e^{-i(\mathbf{kr})}\nabla_{\mathbf{k}}\varphi_{\mathbf{k}}^{*}(\mathbf{r}) = -i\mathbf{r}\psi_{\mathbf{k}}^{*}(\mathbf{r}) + \psi_{\mathbf{k}}^{*}(\mathbf{r})[\nabla_{\mathbf{k}}\ln\varphi_{\mathbf{k}}^{*}(\mathbf{r})], \quad (13.10)$$

or

$$\mathbf{r}\psi_{\mathbf{k}}^{*}(\mathbf{r}) \equiv [i\nabla_{\mathbf{k}} - (i\nabla_{\mathbf{k}}\ln\varphi_{\mathbf{k}}^{*})]\,\psi_{\mathbf{k}}^{*}(\mathbf{r}) \equiv \hat{\mathbf{r}}(\mathbf{k})\,\psi_{\mathbf{k}}^{*}(\mathbf{r}), \quad (13.11)$$

i.e. the operator

$$\hat{\mathbf{r}}(\mathbf{k}) = i\nabla_{\mathbf{k}} - i\,(\nabla_{\mathbf{k}}\ln\varphi_{\mathbf{k}}^{*}) \quad (13.12)$$

is represented by a sum of operators of differentiating with respect to the wave vector (or quasimomentum) and of multiplying by some function of $\mathbf{k}$ (and $\mathbf{r}$). Expanding the functions $\nabla_{\mathbf{k}}\varphi_{\mathbf{k}}^{*}$ in series in functions $\varphi_{\mathbf{k}}$, we may introduce an operator $\hat{\Omega}$ instead of the term $-i\,(\nabla_{\mathbf{k}}\ln\varphi_{\mathbf{k}}^{*})$.

When $U(\mathbf{r}) = \text{const}$ the second term turns zero and $\hat{\mathbf{r}}(\mathbf{k})$ assumes the usual form of the co-ordinate operator in the p-representation, since in this case the quasimomentum coincides with the momentum. Now we can write the expression for the velocity operator in the k-representation:

$$\hat{\mathbf{v}}(\mathbf{k}) = \frac{1}{i\hbar}\{\hat{\mathbf{r}}(\mathbf{k})\,E(\mathbf{k}) - E(\mathbf{k})\,\hat{\mathbf{r}}(\mathbf{k})\} = \frac{1}{\hbar}\frac{dE(\mathbf{k})}{d\mathbf{k}}. \quad (13.13)$$

It takes the *form of an operator of multiplication by the derivative of the energy with respect to quasimomentum:*

$$\hat{\mathbf{v}}(\mathbf{k}) = \frac{1}{\hbar}\frac{dE(\mathbf{k})}{d\mathbf{k}} = \frac{dE}{d\mathbf{P}} = \hat{\mathbf{v}}(\mathbf{P}) = \mathbf{v}. \quad (13.14)$$

This relation is analogous to the expression for the group velocity of a wave packet

$$\mathbf{v}_{gr} = \frac{dE}{d\mathbf{p}}. \quad (13.15)$$

The mean velocity of the electron in the state with energy $E(\mathbf{k})$ (described by the wave function $\psi_{\mathbf{k}}(\mathbf{r})$, but not $\psi_{\mathbf{k}}^{*}(\mathbf{r})$) has a definite value which depends on this state:

$$\langle\mathbf{v}\rangle = \mathbf{v} = \frac{1}{\hbar}\frac{dE}{d\mathbf{k}} = \frac{dE}{d\mathbf{p}} \quad \text{(we leave out the sign } \langle\ \rangle). \quad (13.16)$$

Thus, *the electron velocity averaged over a state with a definite energy* (to be precise, and over an infinitesimal energy interval) *is determined as the derivative of energy with respect to quasimomentum*. At extremal points the average (in the quantum-mecha-

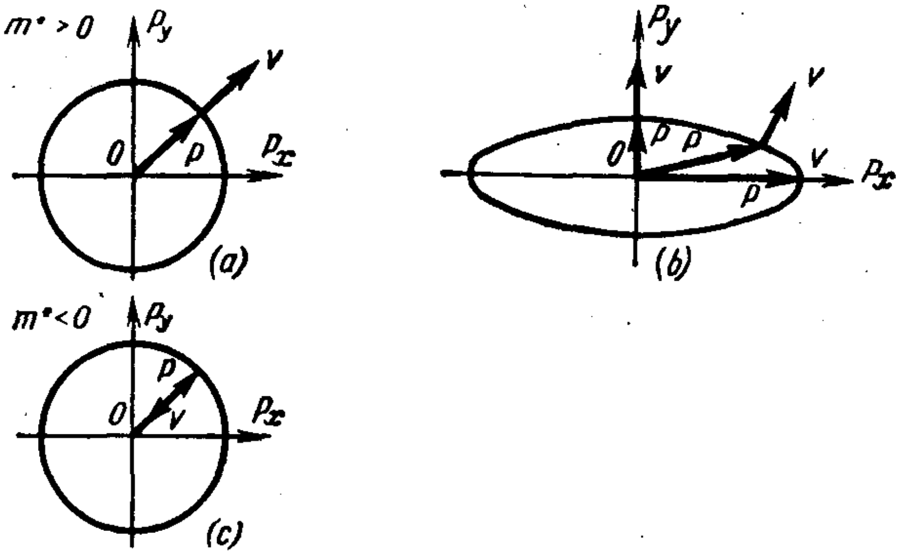


Fig. 9. The direction of the radius vector and of the normal to the energy surface:

(a) spherical energy surfaces; (b) ellipsoidal energy surfaces; (c) spherical energy surfaces, $m^* < 0$

nical sense) velocity is zero. (Below we shall leave out the words "average in the quantum-mechanical sense".)

If we consider states in small elements of volume around extremal points, where the dependence of energy on quasimomentum is quadratic,

$$E-E_0=\frac{1}{2m^*}(\mathbf{P}-\mathbf{P}_0)^2=\frac{1}{2}\sum_{i,j=1}^{3}\frac{(P_i-P_{0i})(P_j-P_{0j})}{m_{ij}}, \quad (13.17)$$

we obtain

$$v_i=\frac{\partial E}{\partial P_i}=\sum_{i=1}^{3}\frac{P_j-P_{0j}}{m_{ij}}, \quad (13.18)$$

or in vector form

$$\mathbf{v}=\frac{1}{\mathbf{m}^*}\cdot(\mathbf{P}-\mathbf{P}_0), \quad (13.19)$$

i.e. *the velocity is generally equal to the scalar product of quasimomentum by the reciprocal effective mass tensor*. If the tensor $\mathbf{m}^{*-1}$ is diagonal

$$\{\mathbf{m}^{*-1}\}_{ij}=m_i^{-1}\delta_{ij}, \quad (13.20)$$

the expression (13.18) will be simplified:

$$v_i=\frac{P_i-P_{0i}}{m_i}. \quad (13.21)$$

Velocity is the energy gradient in the quasimomentum space, therefore it is normal to constant-energy surfaces, $\mathbf{P}-\mathbf{P}_0$ being the radius vector of the points of such a surface. For ellipsoidal energy surfaces the radius vector and the normal are not colinear, and the directions of velocity and quasimomentum do not coincide. They will be colinear only along the directions of ellipsoid axes (Fig. 9), for which case

$$P_i - P_{0i} = \sqrt{2m_i(E-E_0)}, \qquad (13.22)$$

and therefore

$$v_i = \frac{\sqrt{2(E-E_0)}}{\sqrt{m_i}}. \qquad (13.23)$$

Hence, the *velocity along ellipsoid's axes for a given energy is inversely proportional to the square root of the respective effective mass component.*

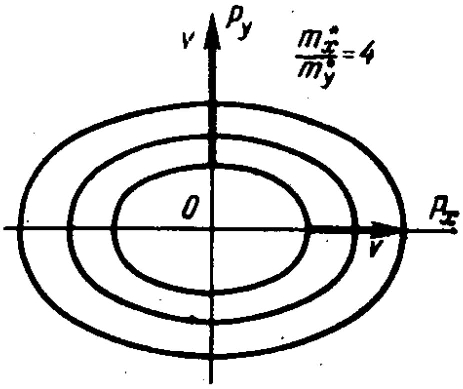


Fig. 10. The relation of the velocity to the magnitude of the effective mass and to the density of energy surfaces

Taking the ellipsoid axes

$$a_i = \sqrt{2m_i(E-E_0)}, \qquad (13.24)$$

we obtain

$$v_i = \frac{a_i}{m_i} = \frac{\sqrt{2(E-E_0)}}{\sqrt{m_i}}, \qquad (13.25)$$

i.e. the more the ellipsoid is elongated, the less is the speed in this direction. This will be obvious if one constructs a family of constant-energy surfaces (Fig. 10). The smaller the value of mass in a given direction, the denser the spacing of constant-energy surfaces and, consequently, the greater the speed along this axis.

Attention should be focused on one essential point related to the sign of the effective mass. Suppose, for the sake of simplicity, the effective mass is a scalar. In this case the vectors v and $(\mathbf{P}-\mathbf{P}_0)$ are colinear; their directions, however, depend on the type of the extremum. For a minimum $m^*>0$ and the directions of velocity and of $(\mathbf{P}-\mathbf{P}_0)$ coincide. For an energy maximum $m^*<0$ and velocity is directed against $(\mathbf{P}-\mathbf{P}_0)$ (see Fig. 9*c*).

14. ACCELERATION OPERATOR

The acceleration operator $\hat{\mathbf{a}}$ *is, by definition, the time derivative of the velocity operator:*

$$\hat{\mathbf{a}} = \frac{d\hat{\mathbf{v}}}{dt} = [\hat{\mathbf{H}},\ \hat{\mathbf{v}}] = \frac{1}{i\hbar}(\hat{\mathbf{v}}\hat{\mathbf{H}} - \hat{\mathbf{H}}\hat{\mathbf{v}}). \qquad (14.1)$$

It is easily seen that for *an electron moving in a periodical lattice field* $\hat{\mathbf{a}} \equiv 0$. Indeed in the **k**-representation $\hat{\mathbf{H}}(\mathbf{k}) \equiv E(\mathbf{k})$ and $\hat{\mathbf{v}}(\mathbf{k}) = \frac{dE(\mathbf{k})}{\hbar d\mathbf{k}}$, therefore

$$\hat{\mathbf{v}}\hat{\mathbf{H}} - \hat{\mathbf{H}}\hat{\mathbf{v}} = 0. \qquad (14.2)$$

But since the commutator of two operators does not depend on the form of representation, it follows that $\hat{\mathbf{a}} \equiv 0$ and

$$\langle\hat{\mathbf{a}}\rangle = \langle\mathbf{a}\rangle = 0. \qquad (14.3)$$

This means that the *electron moves in a periodical field without acceleration*. An intuitive picture of an electron moving with alternate acceleration and deceleration is groundless. This is because the electron possesses the properties of a wave, besides the properties of a particle. This results in the *mean speed being a motion integral and remaining constant* just like the speed of a freely moving particle.

We would like to remind that energy and quasimomentum, too, remain constant in time. The momentum of a particle is, however, not conserved. It changes periodically:

$$\frac{d\mathbf{p}}{dt} = -\nabla U(\mathbf{r}) = \mathbf{F}_l, \qquad (14.4)$$

because the force $\mathbf{F}_l$ with which the lattice acts on the electron has the period of the lattice

$$\mathbf{F}_l(\mathbf{r}+\mathbf{n}) = \mathbf{F}_l(\mathbf{n}). \qquad (14.5)$$

Suppose now an "external" field with potential energy $V(\mathbf{r})$ is applied to the crystal. In this case, as we already know, the quasimomentum will begin to change

$$\frac{d\mathbf{P}}{dt} = \mathbf{F}_a = -\nabla V(\mathbf{r}), \qquad (14.6)$$

provided $V(\mathbf{r}+\mathbf{n}) \neq V(\mathbf{r})$, i.e. the field is not periodic. Quasimomentum changes under the action of an external field $\mathbf{F}_a$ which results from any disturbance of the periodical field. The point in **k**- (or **P**-) space representing the state will move in compliance with the equation

$$\frac{d\mathbf{P}}{dt} = \mathbf{F}_a. \qquad (14.7)$$

From (14.7) we may write

$$\mathbf{P}(t) = \mathbf{P}_0 + \int_0^t \mathbf{F}_a(\xi)\, d\xi. \qquad (14.8)$$

If the external field is time independent, then

$$\mathbf{P}(t) = \mathbf{P}_0 + \mathbf{F}_a t. \tag{14.9}$$

The trajectory of a particle in **P**-space is a straight line which coincides with the direction of the external force $\mathbf{F}_a$. However, as soon as the particle begins to move in the quasimomentum space, it will shift from one energy surface to another. In other words, *external force* $\mathbf{F}_a$ *changes not only the quasimomentum but the particle energy, as well.* The change in energy may be approached from consideration of the Schrödinger equation. This will be done below.

Return now to the acceleration operator. To this end calculate the Poisson's bracket taking into account that an "external" potential field $V(\mathbf{r})$ is applied to the crystal. Denote the lattice field Hamiltonian by $\hat{\mathbf{H}}_0$,

$$\hat{\mathbf{H}}_0 = -\frac{\hbar^2}{2m}\Delta + U(\mathbf{r}). \tag{14.10}$$

For the complete Hamiltonian we have the expression

$$\hat{\mathbf{H}} = -\frac{\hbar^2}{2m}\Delta + U(\mathbf{r}) + V(\mathbf{r}) = \hat{\mathbf{H}}_0 + V(\mathbf{r}). \tag{14.11}$$

Since the velocity operator $\hat{\mathbf{v}}$ commutes with $\hat{\mathbf{H}}_0$, we obtain for the acceleration operator the following expression:

$$\hat{\mathbf{a}} = \frac{d\hat{\mathbf{v}}}{dt} = \frac{1}{i\hbar}\{\hat{\mathbf{v}}(\hat{\mathbf{H}}_0 + \hat{\mathbf{V}}) - (\hat{\mathbf{H}}_0 + \hat{\mathbf{V}})\hat{\mathbf{v}}\} = \frac{1}{i\hbar}(\hat{\mathbf{v}}\hat{\mathbf{V}} - \hat{\mathbf{V}}\hat{\mathbf{v}}). \tag{14.12}$$

Acceleration operator is conveniently expressed in the **k**-representation. Taking into account that

$$\mathbf{v}(\mathbf{k}) = \frac{dE}{\hbar\, dk}, \tag{14.13}$$

we write

$$\hat{\mathbf{a}}(\mathbf{k}) = \frac{1}{i\hbar^2}\left\{\frac{dE}{d\mathbf{k}}\hat{\mathbf{V}}(\mathbf{k}) - \hat{\mathbf{V}}(\mathbf{k})\frac{dE}{d\mathbf{k}}\right\}. \tag{14.14}$$

To calculate $\hat{\mathbf{a}}(\mathbf{k})$ from (14.14) it is necessary to know the potential energy in **k**-representation. To this end **r** in **k**-representation should be substituted into the potential energy $V(\mathbf{r})$ described as a function of the co-ordinate, i.e. the operator

$$\hat{\mathbf{r}}(\mathbf{k}) = i\nabla_{\mathbf{k}} - i(\nabla_{\mathbf{k}} \ln \varphi_{\mathbf{k}}^*) \tag{14.15}$$

should be substituted for **r**.

Consider now the case of a uniform force field most important or practical purposes

$$V(\mathbf{r}) = -(\mathbf{F}_a \mathbf{r}). \qquad (14.16)$$

In the k-representation it will take the form of a sum of an operator of differentiating with respect to **k** and of a multiplication operator:

$$\hat{\mathbf{V}}(\mathbf{k}) = -i(\mathbf{F}_a \nabla_{\mathbf{k}}) + i(\mathbf{F}_a \nabla_{\mathbf{k}} \ln \varphi_{\mathbf{k}}^{*}). \qquad (14.17)$$

Since the second term of (14.17) commutes with $\frac{dE}{\hbar\, d\mathbf{k}}$, it follows that

$$\hat{\mathbf{a}}(\mathbf{k}) = -\frac{i}{\hbar^2}\left\{\frac{dE}{d\mathbf{k}}(-i\mathbf{F}_a\nabla_{\mathbf{k}}) - (-i\mathbf{F}_a\nabla_{\mathbf{k}})\frac{dE}{d\mathbf{k}}\right\} = \frac{1}{\hbar^2}(\mathbf{F}_a\nabla_{\mathbf{k}}^2 E) = \frac{\mathbf{F}_a}{\mathbf{m}^*}. \qquad (14.18)$$

The quantity

$$\frac{1}{\mathbf{m}^*} = \frac{d^2E}{d\mathbf{P}^2} = \frac{1}{\hbar^2}\frac{d^2E}{d\mathbf{k}^2} \qquad (14.19)$$

may be regarded as a generalized reciprocal effective mass tensor which for the case of a quadratic dependence of energy on quasi-momentum coincides with the tensor $\mathbf{m}^{*-1}$ introduced above.

Taking into account (14.19), we write

$$\hat{\mathbf{a}}(\mathbf{k}) = \frac{\mathbf{F}_a}{\mathbf{m}^*(\mathbf{k})}. \qquad (14.20)$$

Since $\mathbf{F}_a$ is independent of **k**, it follows that for an effective mass independent of **k** the form of $\hat{\mathbf{a}}$ will remain the same in any representation:

$$\hat{\mathbf{a}} = \frac{\hat{\mathbf{F}}_a}{\mathbf{m}^*}; \quad \mathbf{a} = \frac{\mathbf{F}_a}{\mathbf{m}^*}, \qquad (14.21)$$

or

$$\mathbf{m}^*\hat{\mathbf{a}} = \hat{\mathbf{F}}_a; \quad \mathbf{m}^*\mathbf{a} = \mathbf{F}_a. \qquad (14.22)$$

Equation (14.21), *or* (14.22), *coincides in its form with Newton's equation of motion.* However, the equation (14.21), or (14.22), has some peculiarities that deserve attention.

To begin with, $\mathbf{m}^*$ is generally a tensor of rank two, and for this reason the *directions of the acceleration vector and the force vector do nót coincide.* Equation (14.21), for example for a_x, takes the form

$$a_x = \frac{1}{m_{xx}}F_x + \frac{1}{m_{xy}}F_y + \frac{1}{m_{xz}}F_z = \sum_{j=1}^{3}\frac{F_j}{m_{xj}}. \qquad (14.23)$$

If m^{*-1} is diagonal, then

$$a_j = \frac{F_j}{m_j}. \tag{14.24}$$

It follows from equation (14.24) that acceleration will be colinear with the force only when the force is directed along one of the ellipsoid axes.

Secondly, *only the external force* $\mathbf{F}_a$ *imparts acceleration to the electron; internal forces* $\mathbf{F}_i$ *cannot impart any acceleration to the electron.*

Thirdly, *the effective mass* $\mathbf{m}^*$ *and not the normal mass serves as a dynamical characteristic of the electron that determines its reaction to external forces.* This means that *though the electron is not accelerated by the lattice field, the latter influences the changes in its motion under external forces.* In other words, *in the presence of external forces the lattice field manifests itself in that the dynamic properties of the electron are no longer determined by its mass, but by its effective mass.*

For a scalar effective mass $\mathbf{a}$ and $\mathbf{F}_a$ are colinear. But the difference between the electron in a lattice and the free electron lies not only in the fact that the value of the effective mass is different from the mass of a free electron. *When the electron is in the vicinity of an energy maximum its effective mass is negative, and its acceleration* $\mathbf{a}$ *is directed against the external force* $\mathbf{F}_a$:

$$\mathbf{a} = \frac{\mathbf{F}_a}{\mathbf{m}^*} = -\frac{\mathbf{F}_a}{|m^*|}. \tag{14.25}$$

Consider a force due to a uniform electric field

$$\mathbf{F}_a = e_n \mathbf{E}, \tag{14.26}$$

where e_n is the electron charge ($e_n < 0$). In this case

$$\mathbf{a} = \frac{e_n}{m^*}\mathbf{E} = \frac{-|e_n|}{-|m^*|}\mathbf{E} = \frac{|e_n|}{|m^*|}\mathbf{E}. \tag{14.27}$$

The acceleration of the electron, Eq. (14.25), in an electric field is in the direction of the field, i.e. *it is directed so as if the electron had a positive charge and a positive effective mass. We will term such an anomalously accelerated electron a quantum-mechanical hole.*

If, on the other hand, the electron is near an energy minimum, then $m^* > 0$, and the acceleration is in the direction of the force, i.e. against the field $\mathbf{E}$, just as in the case of a normal electron.

Expression (14.22) can be derived from (13.19). In fact, differentiating (13.19) with respect to time, we obtain

$$\frac{d\mathbf{v}}{dt} = \frac{d}{dt}\frac{\mathbf{P}-\mathbf{P}_0}{\mathbf{m}^*} = \frac{1}{\mathbf{m}^*}\frac{d\mathbf{P}}{dt} = \frac{1}{\mathbf{m}^*}\mathbf{F}_a = \mathbf{a}. \tag{14.28}$$

Expression (14.28) is identical in form with (14.21); it is, however, more general since in (14.21) $\mathbf{F}_a$ is assumed to be independent of the co-ordinates, while in (14.28) it may be a function of co-ordinates. But, on the other hand, $\mathbf{m}^{*-1}$ in (14.21) is a generalized reciprocal effective mass tensor, while in (14.28) $\mathbf{m}^{*-1}$ is independent of $\mathbf{k}$. In other words, equation (13.19) and, consequently, (14.28) is valid as long as the state $\mathbf{P}-\mathbf{P}_0$ remains inside the area of quadratic dependence of energy on quasimomentum. If the state passes into an area of a more complex dependence of energy on quasimomentum, a time derivative of $\mathbf{m}^{*-1}$ should be added to (14.28).

To wind up the section let us write out the expression for the mean (in a quantum-mechanical sense) momentum of a particle in a stationary state described by a Bloch function $\psi_{\mathbf{k}}(\mathbf{r})$:

$$\langle \mathbf{p}\rangle=\int \psi_{\mathbf{k}}^{*}(\mathbf{r})\,\hat{\mathbf{p}}\psi_{\mathbf{k}}(\mathbf{r})\,d\tau=m\int \psi_{\mathbf{k}}^{*}\frac{\hat{\mathbf{p}}}{m}\psi_{\mathbf{k}}\,d\tau=m\langle \mathbf{v}\rangle. \quad (14.29)$$

But, according to (13.16) and (13.19),

$$\langle \mathbf{v}\rangle=\mathbf{v}=\frac{\mathbf{P}-\mathbf{P}_0}{\mathbf{m}^*}, \quad (14.30)$$

therefore

$$\langle \mathbf{p}\rangle=\frac{m}{\mathbf{m}^*}(\mathbf{P}-\mathbf{P}_0);\quad \mathbf{P}-\mathbf{P}_0\equiv\frac{\mathbf{m}^*}{m}\langle \mathbf{p}\rangle, \quad (14.31)$$

or

$$P_i-P_{0i}=\frac{m_i}{m}\langle p_i\rangle. \quad (14.32)$$

Thus, presuming $\langle \mathbf{p}\rangle$ to have the meaning of a "classical" momentum of the electron in a crystal lattice, we may express the quasimomentum in terms of the "classical" momentum and the effective mass.

Compare the kinetic energy of a free particle with a momentum $\langle \mathbf{p}\rangle$ with the total energy of the electron in a crystal lattice having the same mean momentum:

$$T=\frac{\langle \mathbf{p}\rangle^2}{2m}=\frac{m\mathbf{v}^2}{2}=\frac{m}{2}\left[\frac{(\mathbf{P}-\mathbf{P}_0)}{\mathbf{m}^*}\right]^2. \quad (14.33)$$

For a scalar effective mass $m_i=m^*$ and

$$T=\frac{m}{m^*}\left[E(\mathbf{P})-E(\mathbf{P}_0)\right]. \quad (14.34)$$

Summary of Secs. 12-14

1. The most important concept used to describe the motion of electrons in solids is the effective mass $\mathbf{m}^*$ concept. It is defined as a tensor reciprocal of the reciprocal effective mass tensor $\mathbf{m}^{*-1}$. The

reciprocal effective mass tensor $\mathbf{m}^{*-1}$ is equal to the second derivative of energy with respect to quasimomentum

$$\mathbf{m}^{*-1}=\frac{d^2E}{d\mathbf{P}^2}=\frac{1}{\hbar^2}\frac{d^2E}{d\mathbf{k}^2}, \qquad (14.1s)$$

therefore

$$\mathbf{m}^{*}=\left\{\frac{d^2E}{d\mathbf{P}^2}\right\}^{-1} \qquad (14.2s)$$

If $\mathbf{m}^{*-1}$ is diagonal, $\{\mathbf{m}^{*-1}\}_{ij}=m_i^{-1}\delta_{ij}$, then $\mathbf{m}^*$, too, will be diagonal, and

$$m_{ij}=m_i\delta_{ij}. \qquad (14.3s)$$

If the derivative of energy with respect to $\mathbf{P}$ is calculated at an arbitrary point $\mathbf{P}$, then $\mathbf{m}^*=\mathbf{m}^*(\mathbf{P})$ and $\mathbf{m}^*$ is termed the generalized effective mass.

2. In the vicinity of an extremum $\mathbf{P}_0$ the energy is a quadratic function of $\mathbf{P}-\mathbf{P}_0$ (or $\mathbf{k}-\mathbf{k}_0$) and $\mathbf{m}^{*-1}$, accordingly, does not depend on $\mathbf{P}$. $\mathbf{m}^{*-1}$ is equal to the second derivative at the extremal point. The energy

$$E(\mathbf{P})=E(\mathbf{P}_0)+\frac{1}{2}\frac{(\mathbf{P}-\mathbf{P}_0)^2}{\mathbf{m}^*}=E_0+\sum_{i=1}^{3}\frac{(P_i-P_{0i})^2}{2m_i}. \qquad (14.4s)$$

3. Near an energy minimum the components $m_i>0$, and near an energy maximum $m_i<0$.

4. In the vicinity of an energy extremum constant-energy surfaces are ellipsoids the axes of which are proportional to the square root of the effective mass tensor components. If all such components are equal, effective mass will be a scalar, and constant-energy surfaces will be spherical. If two of the components are equal, for instance $m_1=m_2\neq m_3$, the constant-energy surface will be an ellipsoid of revolution with the axis of rotation corresponding to m_3, which on account of this is termed the longitudinal effective mass and denoted by m_l, while the quantity $m_1=m_2$ is termed the transverse effective mass m_t.

5. The mean (in a quantum-mechanical sense) speed is equal to

$$\langle\mathbf{v}\rangle=\mathbf{v}=\frac{dE}{d\mathbf{P}}=\frac{1}{\hbar}\frac{dE}{d\mathbf{k}}. \qquad (14.5s)$$

Its relation to quasimomentum is

$$\mathbf{v}=\frac{\mathbf{P}-\mathbf{P}_0}{\mathbf{m}^*}. \qquad (14.6s)$$

6. The quasimomentum is related to the mean (in a quantum-mechanical sense) momentum by the formula

$$\langle\mathbf{p}\rangle=m\langle\mathbf{v}\rangle=\frac{m}{\mathbf{m}^*}(\mathbf{P}-\mathbf{P}_0). \qquad (14.7s)$$

7. The electron is accelerated only by an external force $\mathbf{F}_a$. However, the forces of the lattice field $\mathbf{F}_l$ do manifest themselves in that the acceleration is determined by the effective mass $\mathbf{m}^*$:

$$\mathbf{a} = \frac{\mathbf{F}_a}{\mathbf{m}^*} = \mathbf{m}^{*-1}\,\mathbf{F}_a; \quad \mathbf{F}_a = \mathbf{m}^*\mathbf{a}. \tag{14.8s}$$

8. In case of a scalar mass the electron will move against the electric field if it is in the vicinity of an energy minimum. If, on the other hand, it is in the vicinity of an energy maximum its acceleration will be directed against the field. In a given electric field it will move as a particle with a positive effective mass and with a positive charge.

9. Under the influence of an external force $\mathbf{F}_a$ the quasimomentum of the electron will change:

$$\frac{d\mathbf{P}}{dt} = \mathbf{F}_a; \quad \mathbf{P}(t) = \mathbf{P}(0) + \int_0^t \mathbf{F}_a(\xi)\,d\xi. \tag{14.9s}$$

If the force $\mathbf{F}_a$ is independent of time, electron's trajectory in the quasimomentum (and the wave vector) space will be a straight line the direction of which is determined by the direction of the force $\mathbf{F}_a$. As a result of electron motion in $\mathbf{P}$-space its energy changes, and it, under the influence of external forces, passes from one constant-energy surface to another.

10. From (14.6s) and (14.5s) the velocity, or quasimomentum, effective mass, often used in the description of various physical phenomena, may be obtained (for $\mathbf{P}_0 = 0$):

$$\frac{1}{m^*} = \frac{v}{P} = \frac{1}{\hbar k}\cdot\frac{dE}{\hbar\,dk} = \frac{1}{\hbar^2 k}\,\frac{dE}{dk}. \tag{14.10s}$$

For a quadratic dispersion law, different definitions of the effective mass lead to the same value. However, if the dispersion law is other than quadratic, the effective masses derived from different relations will be different.

15. BRILLOUIN ZONES

There is a so-called reciprocal lattice to correspond to every crystal lattice defined as follows.

Construct three vectors $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ using the equations

$$\mathbf{b}_1 = \frac{[\mathbf{a}_2\mathbf{a}_3]}{(\mathbf{a}_1\,[\mathbf{a}_2\mathbf{a}_3])}; \quad \mathbf{b}_2 = \frac{[\mathbf{a}_3\mathbf{a}_1]}{(\mathbf{a}_1\,[\mathbf{a}_2\mathbf{a}_3])}; \quad \mathbf{b}_3 = \frac{[\mathbf{a}_1\mathbf{a}_2]}{(\mathbf{a}_1\,[\mathbf{a}_2\mathbf{a}_3])}. \tag{15.1}$$

With the aid of these vectors we will construct a space lattice. The vectors $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ serve as a basis of this lattice termed reciprocal of the given lattice with a basis $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$.

The volume of an elementary cell of the reciprocal lattice V_b is related by a simple formula to the volume of a direct lattice cell V_a :

$$V_a=(\mathbf{a}_1[\mathbf{a}_2\mathbf{a}_3]);\quad V_b=(\mathbf{b}_1[\mathbf{b}_2\mathbf{b}_3]);\quad V_b=\frac{1}{V_a}. \qquad (15.2)$$

The nodes of the reciprocal lattice are determined by the vector

$$\mathbf{b}=l_1\mathbf{b}_1+l_2\mathbf{b}_2+l_3\mathbf{b}_3, \qquad (15.3)$$

where l_1, l_2, l_3 are integers.

From the definition of the reciprocal lattice basis it follows that

$$(\mathbf{a}_i\mathbf{b}_j)=\frac{1}{V_a}(\mathbf{a}_i[\mathbf{a}_s\mathbf{a}_t])=\delta_{ij}, \qquad (15.4)$$

since when $i=js$ and t cannot coincide with i (or else the scalar product will be equal to unity). When $i\neq j$ one of the vectors $\mathbf{a}_s$ or $\mathbf{a}_t$ coincides with $\mathbf{a}_i$, and the scalar triple product turns zero. This is obvious from the definition of the reciprocal lattice basis.

Consider the scalar product of the vector $\mathbf{b}$ and an arbitrary vector $\mathbf{n}$. Taking into account (15.4), we obtain

$$(\mathbf{nb})=n_1l_1+n_2l_2+n_3l_3=Q, \qquad (15.5)$$

where Q is an integer.

Consider again the translational condition imposed on the wave function of an electron moving in the crystal field:

$$\psi_{\mathbf{k}}(\mathbf{r}+\mathbf{n})=e^{i(\mathbf{kn})}\psi_{\mathbf{k}}(\mathbf{r}). \qquad (15.6)$$

If the vector $\mathbf{k}'=\mathbf{k}+2\pi\mathbf{b}$ is taken instead of the vector $\mathbf{k}$, the translational condition will not be violated

$$e^{i(\mathbf{k}'\mathbf{n})}=e^{i(\mathbf{k}+2\pi\mathbf{b},\,\mathbf{n})}=e^{i(\mathbf{kn})}e^{i2\pi(\mathbf{nb})}=e^{i(\mathbf{kn})}, \qquad (15.7)$$

since

$$(\mathbf{nb})=Q,\quad e^{i2\pi Q}=1 \qquad (15.8)$$

This means that *the states belonging to the vector* $\mathbf{k}$ *and* $\mathbf{k}+2\pi\mathbf{b}$ (*or* $\mathbf{P}$ *and* $\mathbf{P}+2\pi\hbar\mathbf{b}$, *respectively*) *are physically equivalent, and the energy of electrons in these two states should be the same*. In other words, *energy is a periodic function of the wave vector* (or quasi-momentum):

$$E(\mathbf{k}+2\pi\mathbf{b})=E(\mathbf{k}). \qquad (15.9)$$

or

$$E(\mathbf{P}+2\pi\hbar\mathbf{b})=E(\mathbf{P}). \qquad (15.10)$$

If the reciprocal lattice with the basis $2\pi\hbar\mathbf{b}_1$, $2\pi\hbar\mathbf{b}_2$, $2\pi\hbar\mathbf{b}_3$ (or $2\pi\mathbf{b}_1$, $2\pi\mathbf{b}_2$, $2\pi\mathbf{b}_3$) is constructed in $\mathbf{P}$- (or $\mathbf{k}$-) space, the *entire* $\mathbf{P}$- (or $\mathbf{k}$-) *space may be subdivided into areas the sum of whose points represents physically equivalent states. Such areas are termed Bril-*

louin zones. The term the first Brillouin zone refers to a polyhedron of minimum volume built around the origin of co-ordinates of **P**- (or **k**-) *space and containing all possible states.*

Suppose we found such a polyhedron. By adding different vectors $2\pi\hbar\mathbf{b}$ (or $2\pi\mathbf{b}$) to all **P** (or **k**) points of an isolated area we will be able to obtain all the points of **P**- (or **k**-) space. It follows from here that *any points of* **P**- (or **k**-) *space maybe transferred to the main Brillouin zone with the aid of some vector of the reciprocal lattice.* Below we will obtain an equation to enable us to divide **k**-space into Brillouin zones in the most convenient way.

Below we will point out some of the properties of the reciprocal lattice which will be of use to us. Consider the vector **b** with the components

$$\mathbf{b}=(l_1\mathbf{b}_1,\ 0,\ 0). \tag{15.11}$$

It is orthogonal to the plane defined by the vectors $\mathbf{a}_2$ and $\mathbf{a}_3$. The modulus of vector **b** is

$$|\mathbf{b}|=l_1\,|\mathbf{b}_1|=l_1\,\frac{|[\mathbf{a}_2\mathbf{a}_3]|}{V_a}\,. \tag{15.12}$$

For the orthogonal basis of the direct lattice

$$|\mathbf{b}|=l_1\,|\mathbf{b}_1|=\frac{l_1}{a_1}, \tag{15.13}$$

where a_1 is the distance between two neighbouring atomic planes defined by the vectors $\mathbf{a}_2$ and $\mathbf{a}_3$.

This result may be generalized: the vector **b** defines a family of crystal atomic planes orthogonal to it, with the distance d between the planes being related to the reciprocal lattice vector modulus $|\mathbf{b}|$ by the formula

$$|\mathbf{b}|=\frac{l}{d}, \tag{15.14}$$

where l is an integer.

The reciprocal lattice for a cubic crystal is also cubic. If the crystal lattice has a primitive elementary cell, so, too, has the reciprocal lattice. The *Brillouin zone in* **k**-*space for a crystal with a simple cubic lattice is a cube of volume* $\frac{8\pi^3}{a^3}$. To prove it construct a lattice in **k**-space with the basis $2\pi\mathbf{b}_1$, $2\pi\mathbf{b}_2$, $2\pi\mathbf{b}_3$. The cube built on these vectors contains non-equivalent points since these points cannot be transformed one into the other with the aid of some vector **b**. This does not apply to points lying on the faces of the cube and spanned by vectors $2\pi\mathbf{b}_i$ or $-2\pi\mathbf{b}_i$. All points lying outside the cube may be obtained as projections of points lying inside the cube. This fact enables us to say that the isolated volume contains all physically non-equivalent states, i.e. constitutes a Brillouin zone. To construct the first Brillouin zone all points should be displaced by the vector $(-\pi \mathbf{b}_1, -\pi \mathbf{b}_2, -\pi \mathbf{b}_3)$; as a result the centre of the cube will be made to coincide with the origin $\mathbf{k}=0$.

Hence, *the values of the vector* **k** *lie within the range*

$$\begin{aligned} -\frac{\pi}{a} &\leqslant k_x < \frac{\pi}{a}, \\ -\frac{\pi}{a} &\leqslant k_y < \frac{\pi}{a}, \\ -\frac{\pi}{a} &\leqslant k_z < \frac{\pi}{a}. \end{aligned} \tag{15.15}$$

If the crystal lattice is not a cubic one, we may write for each component k_i

$$-\frac{\pi}{a_i} \leqslant k_i < \frac{\pi}{a_i} \tag{15.16}$$

or, in a general form,

$$-\frac{\pi}{a_i} + 2\pi l_i b_i \leqslant k_i < \frac{\pi}{a_i} + 2\pi l_i b_i, \tag{15.17}$$

where l_i is an arbitrary integer.

It follows from the equivalence of points belonging to different Brillouin zones that *the path of a particle moving in* **k**- (or **P**-) *space may be considered as confined to the first Brillouin zone.* To this end the states corresponding to some point of the boundary should be transferred to an equivalent point of the opposite boundary of the Brillouin zone, as is shown in Fig. 11. We will demonstrate now that the reciprocal lattice of a cubic lattice having a volume-centred (or a face-centred) elementary cell is cubic with a face-centred (or a volume-centred) elementary cell.

In a volume-centred cubic lattice (VCC) there are eight sites in the apexes of the cube and one in its centre. Each apex site belongs to eight cells, therefore there are $8\times 1/8+1=2$ sites to a cell. The volume of the cell is a^3, or $a^3/2$ per site. The apex co-ordinates may be written in the form $a\,(0, 0, 0)$; $a\,(1, 0, 0)$; $a\,(0, 1, 0)$; $a\,(0, 0, 1)$; $a\,(1, 1, 1)$; $a\,(0, 1, 1)$; $a\,(1, 0, 1)$ and $a\,(1, 1, 0)$ and the centre co-ordinate $a\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$.

Choose three vectors of the form

$$\left.\begin{aligned} \mathbf{a}_1 &= \frac{a}{2}(\mathbf{i}+\mathbf{j}+\mathbf{k}), \\ \mathbf{a}_2 &= \frac{a}{2}(-\mathbf{i}+\mathbf{j}+\mathbf{k}), \\ \mathbf{a}_3 &= \frac{a}{2}(-\mathbf{i}-\mathbf{j}+\mathbf{k}), \end{aligned}\right\} \tag{15.18}$$

i.e. three vectors from the origin to the centres of three nearest elementary cells. The co-ordinate axes are determincd by unit vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ directed along the edges of the elementary cell. It is readily seen that any lattice site may be obtained with the aid of the translation vector $\mathbf{n} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$. After some simple transformations the volume of the basis parallelepiped is calculated to be $V_a = a^3/2$, i.e, equal to the volume per site.

For a face-centred cubic lattice with the period a there are eight sites in the apexes of the cell and six sites in the centres of the faces. Altogether there are $8\times 1/8 + 6\times 1/2 = 4$ sites per elementary cell, or the volume $a^3/4$ per site. Three vectors from the site at the origin to the sites on the face centres may be taken for the basis vectors, for instance,

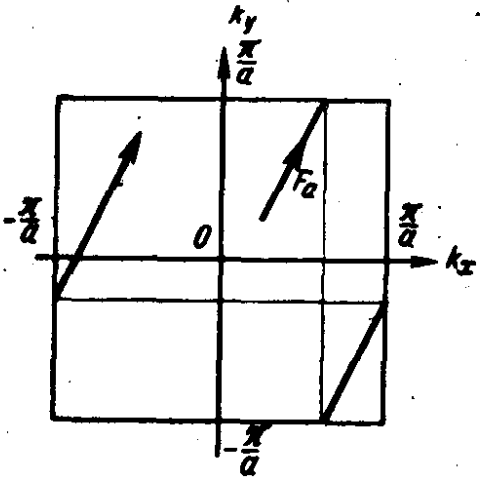


Fig. 11. The motion of the electron inside the main Brillouin zone

$$\left.\begin{aligned}\mathbf{a}_1 &= \frac{a}{2}(\mathbf{i}+\mathbf{j}),\\ \mathbf{a}_2 &= \frac{a}{2}(\mathbf{j}+\mathbf{k}),\\ \mathbf{a}_3 &= \frac{a}{2}(\mathbf{k}+\mathbf{i}).\end{aligned}\right\} \qquad (15.19)$$

The volume of the basis parallelepiped is $V_a = (\mathbf{a}_1[\mathbf{a}_2\mathbf{a}_3]) = a^3/4$.

Construct now the basis of the reciprocal lattice. For the FCC-lattice we obtain, in accordance with (15.1) and (15.19),

$$\mathbf{b}_1 = \frac{1}{V_a}[\mathbf{a}_2\mathbf{a}_3] = \frac{1}{\frac{a^3}{4}}\cdot\frac{a^2}{4}[\mathbf{j}+\mathbf{k},\ \mathbf{k}+\mathbf{i}] = \frac{1}{a}(\mathbf{i}-\mathbf{k}+\mathbf{j}) = \frac{b}{2}(\mathbf{i}+\mathbf{j}-\mathbf{k});$$

$$\begin{aligned}\mathbf{b}_2 &= \frac{b}{2}(\mathbf{j}+\mathbf{k}-\mathbf{i});\\ \mathbf{b}_3 &= \frac{b}{2}(\mathbf{k}+\mathbf{i}-\mathbf{j}).\end{aligned} \qquad (15.20)$$

Hence, if we put the period of the reciprocal lattice equal to $b = \frac{2}{a}$, the volume of its elementary cell will be $8/a^3$, and the reciprocal lattice itself will be of the VCC type.

For the VCC-lattice we find, in compliance with (15.1) and (15.18),

$$\mathbf{b}_1 = \frac{1}{V_a}[\mathbf{a}_2\mathbf{a}_3] = \frac{1}{\frac{a^3}{2}} \cdot \frac{a^2}{4}[-\mathbf{i}+\mathbf{j}+\mathbf{k},\ -\mathbf{i}-\mathbf{j}+\mathbf{k}] =$$

$$= \frac{1}{a}(\mathbf{i}+\mathbf{k}) = \frac{b}{2}(\mathbf{i}+\mathbf{k});$$

$$\mathbf{b}_2 = \frac{b}{2}(-\mathbf{i}+\mathbf{j});$$

$$\mathbf{b}_3 = \frac{b}{2}(-\mathbf{j}+\mathbf{k}); \quad b = \frac{2}{a}. \tag{15.21}$$

As may be seen from (15.20) and (15.21), the reciprocal lattice for a VCC-lattice is of the FCC type, and vice versa. The relation between the direct and the reciprocal lattice periods is $ab = 2$ and that between the volumes—$V_a V_b = 1$.

The reciprocal lattice for a hexagonal lattice with $c/a > 1$ is a hexagonal lattice contracted along the hexagonal axis.

The above may be taken as a proof that the symmetry of the direct and reciprocal lattices is similar, therefore, it may be expected that the symmetry of the $E(\mathbf{k})$ dependence is completely established by the symmetry of the lattice potential field $U(\mathbf{r})$.

16. NORMALIZING INSIDE A POTENTIAL BOX AND THE DISCRETE NATURE OF QUASIMOMENTUM

In all calculations carried out in quantum mechanics the wave function is usually assumed to be normalized.

The process of normalizing the eigenfunctions of an operator depends on its eigenvalues spectrum. For a discrete spectrum the wave functions are normalized to unity:

$$\int \psi_n^*(\mathbf{r})\,\psi_m(\mathbf{r})\,d\tau = \delta_{mn} = \begin{cases} 1 \text{ at } n=m, \\ 0 \text{ at } n \neq m, \end{cases} \tag{16.1}$$

where $\psi_n(\mathbf{r})$, $\psi_m(\mathbf{r})$ are the eigenfunctions (in $\mathbf{r}$-representation) of some operator $\hat{M}$ corresponding to the eigenvalues M_m and M_n:

$$\hat{M}\psi_n(\mathbf{r}) = M_n\psi_n(\mathbf{r}); \quad \hat{M}\psi_m(\mathbf{r}) = M_m\psi_m(\mathbf{r}). \tag{16.2}$$

If the eigenvalues spectrum of the operator $\hat{M}$ is continuous,

$$\hat{M}\psi_M(\mathbf{r}) = M\psi_M(\mathbf{r}); \quad \hat{M}\psi_{M'}(\mathbf{r}) = M'\psi_{M'}(\mathbf{r}), \tag{16.3}$$

then the eigenfunctions are normalized to the Dirac δ-function,

$$\int \psi_M^*(\mathbf{r})\,\psi_{M'}(\mathbf{r})\,d\tau = \delta(M-M'), \tag{16.4}$$

where the Dirac δ-function satisfies the well-known conditions

$$\delta(x-a)=\begin{cases}0 & \text{at } x\neq a,\\ \infty & \text{at } x=a,\end{cases}$$

$$\int_b^c \delta(x-a)\,dx=\begin{cases}1 & \text{at } a\in[b,c]\\ 0 & \text{at } a\notin[b,c]\end{cases} \quad (16.5)$$

In the process of normalizing the integration should be performed over the entire volume where the particle may be found.

In reality pa ticles move in a confined space, for instance, in a finite crystal. In this case it will be necessary to solve the Schrödinger equation for a finite crystal, and this requires the knowledge of the boundary conditions, i.e. the values of the wave function itself and its first derivatives with respect to co-ordinates on the crystal boundary. This leads to quite a number of additional complications, i.e. the crystal's translational symmetry is disturbed, additional boundary conditions have to be introduced which are not easy to take account of since crystal surface phenomena are probably more involved than volume phenomena. It is at the same time obvious that if the crystal dimensions are sufficiently large, surface phenomena will not influence decisively the processes in the bulk of the crystal. This fact enables the so-called *cyclic boundary conditions* to be introduced.

Suppose the crystal is of the shape of a parallelepiped with sides L_1, L_2, L_3 and volume $G=L_1L_2L_3$.

Suppose the entire space is filled with such crystals. In this case the translational property of the crystal field is retained. Since all the points spaced at several L_1, L_2, L_3 are equivalent, we have the condition of equivalence of physical properties of the crystal at points x and $x+L_1$ and at analogous points, instead of the usual boundary conditions. Accordingly, the *place of usual boundary conditions is taken by the cyclic boundary conditions*

$$\psi(x, y, z)=\psi(x+L_1, y, z)=\psi(x, y+L_2, z)= \\ =\psi(x, y, z+L_3). \quad (16.6)$$

In this case electron motion may be considered only inside the main area G which my be taken as the normalizing area:

$$\int_G \psi_M^*(\mathbf{r})\,\psi_{M'}(\mathbf{r})\,d\tau=1. \quad (16.7)$$

This so-called "normalizing inside a box" results in certain changes of the spectrum of M which will be considered for the case of quasimomentum eigenvalues. As we have seen, quasimomentum is subject to one condition—its value should be real.

Since it may prove to be continuous, its eigenfunctions, i.e. Bloch functions, should be normalized to δ-function:

$$\int \psi_{\mathbf{k}}(\mathbf{r})\,\psi^{*}_{\mathbf{k}'}(\mathbf{r})\,d\tau=\delta(\mathbf{k}-\mathbf{k}'). \tag{16.8}$$

The integration here is carried out over the entire space. Normalizing inside a box, we may write

$$\int \psi^{*}_{\mathbf{k}}(\mathbf{r})\,\psi_{\mathbf{k}'}(\mathbf{r})\,d\tau=\delta_{\mathbf{k}\mathbf{k}'}=\begin{cases}1 \text{ at } \mathbf{k}=\mathbf{k}'\\ 0 \text{ at } \mathbf{k}\neq\mathbf{k}'.\end{cases} \tag{16.9}$$

In other words, we write the normalization condition for an operator function with a discrete spectrum. A discrete spectrum, however, follows directly from the cyclic boundary conditions:

$$\begin{aligned}\psi_{\mathbf{k}}(x+L_1,\,y,\,z)&=e^{i[k_x(x+L_1)+k_y y+k_z z]}\,\varphi_{\mathbf{k}}(x+L_1,\,y,\,z)=\\ &=e^{ik_xL_1}e^{i(\mathbf{kr})}\,\varphi_{\mathbf{k}}(x+L_1,\,y,\,z)=e^{i(\mathbf{kr})}\,\varphi_{\mathbf{k}}(\mathbf{r}).\end{aligned} \tag{16.10}$$

The condition $\varphi_{\mathbf{k}}(x+L_1,\,y,\,z)=\varphi_{\mathbf{k}}(x,\,y,\,z)$ is satisfied automatically since L_1, L_2, L_3 contain integral numbers of lattice periods and $\varphi_{\mathbf{k}}(\mathbf{r})$ is periodic with the period of the direct lattice. It follows from the last equation that

$$e^{ik_xL_1}=1. \tag{16.11}$$

This is possible only if the *power of the imaginary exponent is an integer* multiplied by 2π, i.e.

$$k_xL_1=2\pi n_1, \tag{16.12}$$

where n_1 is an arbitrary integer. It follows from here that the wave vector is discrete

$$k_x=\frac{2\pi}{L_1}n_1,\quad n_1=0,\ \pm1,\ \pm2,\ \ldots. \tag{16.13}$$

With due regard to the conditions of periodicity along the axes Oy and Oz we obtain likewise

$$\left.\begin{aligned}k_y&=\frac{2\pi}{L_2}n_2,\quad n_2=0,\ \pm1,\ \pm2,\ \ldots\\ k_z&=\frac{2\pi}{L_3}n_3,\quad n_3=0,\ \pm1,\ \pm2,\ \ldots\end{aligned}\right\} \tag{16.14}$$

Define L_i in terms of a number N_i of lattice sites arranged along the edge of the crystal

$$L_i=N_ia_i\quad(i=1,\,2,\,3). \tag{16.15}$$

Substituting L_i into the expression for k_i, we obtain:

$$k_i=\frac{2\pi}{a_i}\frac{n_i}{N_i}=2\pi b_i\frac{n_i}{N_i}. \tag{16.16}$$

Taking into account the equivalence of the states with $\mathbf{k}$ and $\mathbf{k}'=\mathbf{k}+2\pi\mathbf{b}$, we can bound n_i above by the condition

$$k_i=2\pi b_i,\quad n_i=N_i, \tag{16.17}$$

with $n_i=0$ as the lower limit.

There is also the possibility of choosing another condition, $\frac{n_i}{N_i}=\pm\frac{1}{2}$, which results in a range of values symmetrical with respect to $\mathbf{k}=0$:

$$|k_i|\leqslant\frac{1}{2}2\pi b_i=\frac{\pi}{a_i} \tag{16.18}$$

or

$$-\frac{\pi}{a_i}\leqslant k_i<\frac{\pi}{a_i},\quad n_i=0,\ \pm1,\ \pm2,\ \ldots,\ \pm\frac{N_i-1}{2},\ \pm\frac{N_i}{2}. \tag{16.19}$$

Thus, for a *finite crystal the wave vector assumes discrete values*. However, this discreteness is not essential for a crystal of sufficiently large size, and we will assume the quantity k_i *to be quasicontinuous*. However, there are cases when such discreteness should be taken into account. Since the quasimomentum assumes discrete values so does the energy $E(\mathbf{k}_i)$. We shall in future disregard this fact because the distance between energy levels is much smaller than kT.

Let us return to the conditions of normalizing a wave function inside a "box". Substituting the Bloch function into (16.9)

$$\int e^{-i(\mathbf{k}'\mathbf{r})}\varphi^*_{\mathbf{k}'}(\mathbf{r})e^{i(\mathbf{kr})}\varphi_{\mathbf{k}}(\mathbf{r})\,d\tau=\delta_{\mathbf{kk}'}, \tag{16.20}$$

we obtain for $k=k'$

$$\int\varphi^*_{\mathbf{k}}(\mathbf{r})\,\varphi_{\mathbf{k}}(\mathbf{r})\,d\tau=1. \tag{16.21}$$

It follows from (16.21) that the integral of $|\varphi_{\mathbf{k}}(\mathbf{r})|^2$ over the area G should be unity. This condition determines the choice of the multiplicative constant in $\varphi_{\mathbf{k}}(\mathbf{r})$.

Since, however, $\varphi_{\mathbf{k}}(\mathbf{r})$ is a periodic function, it suffices to perform integration of $|\varphi_{\mathbf{k}}(r)|^2$ over the volume $V_a=(\mathbf{a}_1[\mathbf{a}_2\mathbf{a}_3])$ of one cell. If the number of such cells in a crystal is N, then evidently

$$\int_G|\varphi_{\mathbf{k}}(\mathbf{r})|^2\,d\tau=N\int_{V_a}|\varphi_{\mathbf{k}}(\mathbf{r})|^2\,d\tau=1, \tag{16.22}$$

which enables us to write the normalizing condition for $\varphi_{\mathbf{k}}(\mathbf{r})$ in the form

$$\int_{V_a}|\varphi_{\mathbf{k}}(\mathbf{r})|^2\,d\tau=N^{-1}. \tag{16.23}$$

Traditionally, however, $\varphi_{\mathbf{k}}(\mathbf{r})$ is normalized to unity over the volume V_a:

$$\int\limits_{V_a} |\varphi_{\mathbf{k}}(\mathbf{r})|^2 d\tau = 1. \qquad (16.24)$$

In this case $\psi(\mathbf{r})$ normalized to unity should be of the form

$$\psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{N}} e^{i(\mathbf{kr})} \varphi_{\mathbf{k}}(\mathbf{r}). \qquad (16.25)$$

Summary of Secs. 15-16

1. There is a so-called reciprocal lattice defined by its basis $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ to correspond to every lattice with the basis $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$.

2. The states $\mathbf{k}$ and $\mathbf{k}' = \mathbf{k} + 2\pi\mathbf{b}$ are physically equivalent and because of this energy is a periodic function of the wave vector (and quasimomentum)

$$E(\mathbf{k} + 2\pi\mathbf{b}) = E(\mathbf{k}). \qquad (16.1s)$$

The $\mathbf{k}$- (or $\mathbf{P}$-) space is divided into equivalent areas each of which contains all possible states. These areas are termed Brillouin zones. The zone symmetrical with respect to the origin of co-ordinates is termed basic Brillouin zone. It determines the set of possible values of the wave vector:

$$-\frac{\pi}{a_i} \leqslant k_i < \frac{\pi}{a_i}. \qquad (16.2s)$$

3. In order to exclude boundary conditions of the Schrödinger equation for a finite crystal without loss of translational symmetry it is assumed that the entire space is filled with crystals having the shape of parallelepipeds with the edges L_1, L_2, L_3. In this case states at points spaced at intervals multiple to L_i along each axis are presumed to be equivalent, and for this reason the wave function is normalized to unity over the volume $G = L_1L_2L_3$ (over the volume of the "box").

4. It follows from the boundary conditions that the wave vector (and quasimomentum) are discrete quantities. The difference between adjacent values $\Delta k_i = k_{i,\,n+1} - k_{i,\,n}$ is equal to

$$\Delta k_i = \pm \frac{2\pi}{L_i}. \qquad (16.3s)$$

5. The discrete nature of the wave vector results in discrete energy surfaces in the Brillouin zone. They set up numerous energy sublevels (or levels). However, since Δk_i is small (k_i being a quasicontinuous quantity) $E(\mathbf{k})$, too, is a quasicontinuous quantity.

17. THEORY OF THE QUASIFREE ELECTRON

In treating electron motion in a periodic field we derived motion integrals, introduced the concept of effective mass and established the relationship between (mean) velocity, quasimomentum, acceleration and external force. All these quantities are related directly, or indirectly, to the dependence of energy on quasimomentum $E(\mathbf{P})$, or $E(\mathbf{k})$. As was already mentioned, the revelation of the dependence $E(\mathbf{k})$ in a general form constitutes an important problem of solid state physics as yet unsolved. Approximate methods of establishing the general nature of $E(\mathbf{k})$ dependence involving in some cases the use of experimental quantities needed for the practical application of theoretical calculations are, however, of great value for understanding many processes in solids, particularly in semiconductors.

To find the generalized dependence $E(\mathbf{k})$ let us resort to the perturbation theory method. In solving the problem of the elect ron moving in a periodic field there can be *two approximations distinguished by the choice of the zero approximation. With the free particle as the zero approximation and the periodic field as a perturbation we arrive at the so-called theory of the quasifree electron. If the electron in an isolated atom is taken as zero approximation we will arrive at the quasibound electron theory.* Consider the former theory.

Represent the periodic field Hamiltonian

$$\hat{\mathbf{H}} = \hat{\mathbf{T}} + \hat{\mathbf{U}}(\mathbf{r}) \tag{17.1}$$

as the sum of the "unperturbed" system Hamiltonian $\hat{\mathbf{H}}_0 = \hat{\mathbf{T}}$ and the "perturbation" $\hat{\mathbf{W}}(\mathbf{r}) = \hat{\mathbf{U}}(\mathbf{r})$:

$$\hat{\mathbf{H}} = \hat{\mathbf{H}}_0 + \hat{\mathbf{W}}(\mathbf{r}) = \hat{\mathbf{T}} + \hat{\mathbf{U}}(\mathbf{r}),\ \hat{\mathbf{U}}(\mathbf{r}) \equiv U(\mathbf{r}) \tag{17.2}$$

The task of the perturbation theory is to find corrections to the energy E^0 and to the wave function ψ^0 of the unperturbed system when perturbation is applied. Determine the energy spectrum and wave functions of a free particle.

The solution of the Schrödinger equation for a free particle, i.e., for a particle with the Hamiltonian $\hat{\mathbf{H}}_0 = \hat{\mathbf{T}}$

$$-\frac{\hbar^2}{2m}\Delta\psi^0(\mathbf{r}) = E^0\psi^0(\mathbf{r}) \tag{17.3}$$

takes the form of plane de Broglie waves

$$\psi^0_{\mathbf{k}}(\mathbf{r}) = Ae^{i(\mathbf{kr})} \tag{17.4}$$

with a continuous energy spectrum

$$E^0(\mathbf{k}) = \frac{\hbar^2 \mathbf{k}^2}{2m} = \frac{\mathbf{p}^2}{2m}. \tag{17.5}$$

Thus in zero approximation electron energy in the crystal is continuous and has a quadratic dependence on the wave vector and quasimomentum which in this case coincides with the ordinary momentum. The effective mass is a zero rank tensor (scalar) identical to the ordinary free electron mass:

$$m^{*-1} = \frac{d^2E^0}{\hbar^2\, d\mathbf{k}^2} = m^{-1}, \quad m^* \equiv m. \tag{17.6}$$

The amplitude A of the wave $\psi_{\mathbf{k}}^0(\mathbf{r})$ is $\frac{1}{\sqrt{(2\pi)^3}}$ when normalized to δ-function and integrated over the entire space, or $\frac{1}{\sqrt{G}}$ when normalized to unity and integrated over the area G. Hence, the normalized eigenfunction assumes the form

$$\psi_{\mathbf{k}}^0(\mathbf{r}) = \frac{1}{(2\pi)^{\frac{3}{2}}} e^{i(\mathbf{kr})}; \quad \int\limits_{(\infty)} \psi_{\mathbf{k}'}^{0*}(\mathbf{r})\, \psi_{\mathbf{k}}^0(\mathbf{r})\, d\tau = \delta(\mathbf{k}-\mathbf{k}'); \tag{17.7}$$

$$\psi_{\mathbf{k}}^0(\mathbf{r}) = \frac{1}{G^{\frac{1}{2}}} e^{i(\mathbf{kr})}; \quad \int\limits_{G} \psi_{\mathbf{k}'}^{0*}(\mathbf{r})\, \psi_{\mathbf{k}}^0(\mathbf{r})\, d\tau = \delta_{\mathbf{kk}'}. \tag{17.8}$$

To find energy and eigenfunctions in the first approximation we will need, in accordance with the perturbation theory, the matrix elements of perturbation:

$$\mathbf{W}_{\mathbf{k}'\mathbf{k}} = \mathbf{U}_{\mathbf{k}'\mathbf{k}} = \int \psi_{\mathbf{k}'}^{0*}(\mathbf{r})\, U(\mathbf{r})\, \psi_{\mathbf{k}}^0(\mathbf{r})\, d\tau. \tag{17.9}$$

It is easily shown that the matrix elements are non-zero only for definite values of $\mathbf{k}'$ and $\mathbf{k}$. To prove it expand $U(\mathbf{r})$ into a triple Fourier series (this is possible since $U(\mathbf{r})$ is periodic with the periods $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$):

$$U(\mathbf{r}) = \sum_{l_1 l_2 l_3 = -\infty}^{\infty} c_{l_1 l_2 l_3} e^{i2\pi\left(\frac{l_1 x}{a_1} + \frac{l_2 y}{a_2} + \frac{l_3 z}{a_3}\right)} = \sum_{b} c_{\mathbf{b}} e^{i2\pi(\mathbf{br})}, \tag{17.10}$$

$$\mathbf{b} = l_1\mathbf{b}_1 + l_2\mathbf{b}_2 + l_3\mathbf{b}_3. \tag{17.11}$$

Substituting (17.10) into (17.9) we obtain

$$U_{\mathbf{k}'\mathbf{k}} = \frac{1}{G}\int\limits_{G} e^{-i(\mathbf{k}'\mathbf{r})} \sum_{\mathbf{b}} c_{\mathbf{b}} e^{i2\pi(\mathbf{br})} e^{i(\mathbf{kr})}\, d\tau = \frac{1}{G}\sum_{\mathbf{b}} c_{\mathbf{b}} \int\limits_{G} e^{i(\mathbf{k}+2\pi\mathbf{b}-\mathbf{k}',\, \mathbf{r})}\, d\tau =$$

$$= \sum_{\mathbf{b}} c_{\mathbf{b}} \delta_{\mathbf{k}',\, \mathbf{k}+2\pi\mathbf{b}} = \begin{cases} 0, & \mathbf{k}' \neq \mathbf{k} + 2\pi\mathbf{b} \\ c_{\mathbf{b}}, & \mathbf{k}' = \mathbf{k} + 2\pi\mathbf{b}. \end{cases} \tag{17.12}$$

Thus, the *matrix elements* $U_{\mathbf{k}'\mathbf{k}}$ are *equal either to zero* or *to the Fourier series coefficients of* U (**r**).

The *correction to the energy* E^0 (**k**) *in the first approximation* E' (**k**) *is equal to the diagonal matrix element of the perturbation operator:*

$$E'(\mathbf{k}) = U_{\mathbf{kk}} = \frac{1}{G}\int_G U(\mathbf{r})\,d\tau = \langle U\rangle \qquad (17.13)$$

i.e. E'(**k**) *is equal to the value of the potential energy averagεd over the area G*; this correction is independent of **k**.

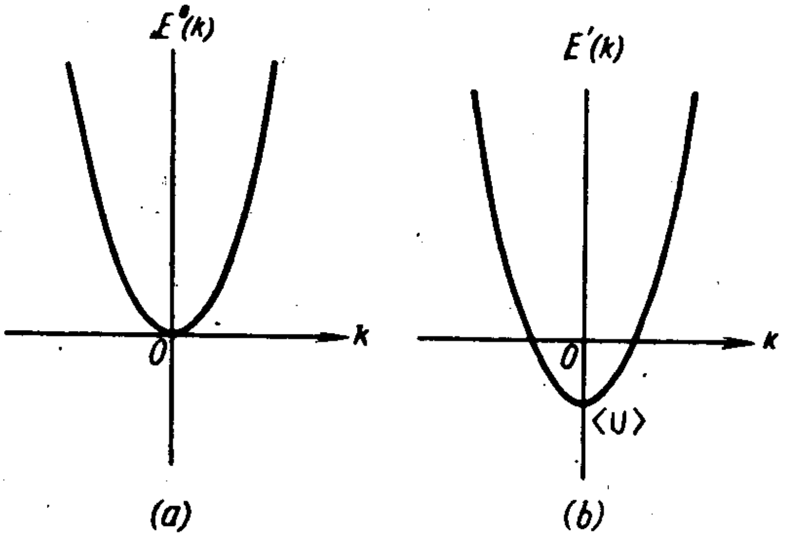


Fig. 12. The energy spectrum of the electron in the theory of the quasifree electron in the zero (*a*) and in the first (*b*) approximation with respect to perturbation

We see from here that in the first approximation of the perturbation theory the only change in the energy spectrum caused by the application of a periodic field is a uniform shift of all the energy levels E^0 (**k**) by the same value $\langle U\rangle$:

$$E^{(1)}(\mathbf{k}) = E^0(\mathbf{k}) + E'(\mathbf{k}) = \frac{\hbar^2\mathbf{k}^2}{2m} + \langle U\rangle. \qquad (17.14)$$

The value $\langle U\rangle$ corresponds to the depth of the potential trough serving as a model of the real crystal in the Sommerfeld elementary theory of metals. $\langle U\rangle$ is the true work function of the electron in the solid. Figure 12 shows the electron spectrum in the zero and the first approximation. Thus, there is no change in the energy spectrum except the displacement of the energy reference point. Choose it so that $\langle U\rangle = 0$.

Consider the correction to the energy in the second approximation. The value of the energy in the second approximation may be written as follows:

$$E^{(2)}(\mathbf{k})=E^0(\mathbf{k})+E'(\mathbf{k})+E''(\mathbf{k})=E^0(\mathbf{k})+\langle U\rangle+$$
$$+\sum_{\mathbf{k}'}\frac{|U_{\mathbf{k}'\mathbf{k}}|^2}{E^0(\mathbf{k})-E^0(\mathbf{k}')}=E^0(\mathbf{k})+\sum_{\mathbf{b}}\frac{|c_{\mathbf{b}}|^2}{E^0(\mathbf{k})-E^0(\mathbf{k}+2\pi\mathbf{b})}. \quad (17.15)$$

We see that *in the second approximation of the perturbation theory the energy correction is proportional to* $|c_{\mathbf{b}}|^2$. *Therefore, when*

$$E^0(\mathbf{k})\gtrless E^0(\mathbf{k}+2\pi\mathbf{b}), \quad (17.16)$$

the energy correction is negligible, and the energy spectrum does not change. In this case the wave function takes the form:

$$\psi_{\mathbf{k}}^{(1)}(\mathbf{r})=\psi_{\mathbf{k}}^0(\mathbf{r})+\sum_{\mathbf{b}}\frac{c_{\mathbf{b}}}{E^0(\mathbf{k})-E^0(\mathbf{k}+2\pi\mathbf{b})}\psi_{\mathbf{k}+2\pi\mathbf{b}}^0(\mathbf{r}), \quad (17.17)$$

and under identical conditions we may contend that perturbation introduces small changes into the wave function which describes free motion.

However, the states should be considered for which the energy difference $E^0(\mathbf{k})-E^0(\mathbf{k}+2\pi\mathbf{b})$ is comparable with $|c_{\mathbf{b}}|^2$. In this case the wave function endures a great change since the free motion of the electron is greatly perturbed. This perturbation will be maximal when the denominator for some state turns zero. Suppose, for example, some term of the sum, say for $\mathbf{b}=\mathbf{g}$, has a denominator close, or equal, to zero. The expressions (17.15) and (17.17) for energy and wave function in this case become meaningless since the theory of perturbations can no more be applied. *When* $E^0(\mathbf{k})\to E^0(\mathbf{k}+2\pi\mathbf{g})$ *the coefficient with* $\psi_{\mathbf{k}+2\pi\mathbf{g}}^0(\mathbf{r})$ *tends towards* ∞. *This means that the share of* $\psi_{\mathbf{k}+2\pi\mathbf{g}}^0(\mathbf{r})$ *in the state* $\psi_{\mathbf{k}}^{(1)}(\mathbf{r})$ *is no less than the part played by* $\psi_{\mathbf{k}}^0(\mathbf{r})$. *When* $E^0(\mathbf{k})=E^0(\mathbf{k}+2\pi\mathbf{g})$ *precisely, the state* $E^0(\mathbf{k})$ *is degenerate* since now two different functions $\psi_{\mathbf{k}}^0(\mathbf{r})$ and $\varphi_{\mathbf{k}+2\pi\mathbf{g}}^0(\mathbf{r})$ correspond to the same value of energy. But this means that for the wave function $\psi^0(\mathbf{r})$, already in the zero approximation, both wave functions should be taken into account, and the solution of the problem should account for this degeneracy. The zero approximation of the wave function $\psi^0(\mathbf{r})$ should, in compliance with the perturbation theory, take the form:

$$\psi^0(\mathbf{r})=\alpha\psi_{\mathbf{k}}^0(\mathbf{r})+\beta\psi_{\mathbf{k}+2\pi\mathbf{g}}^0(\mathbf{r}) \quad (17.18)$$

where α and β are unknown coefficients. For non-degenerate states $\beta\ll\alpha$, for degenerate states α and β are of the same order of magnitude. This is true not only when the energies are precisely equal but whenever

$$|E^0(\mathbf{k})-E^0(\mathbf{k}+2\pi\mathbf{g})|\leqslant|c_g|^2. \tag{17.19}$$

Thus, choosing the function (17.18) for the zero approximation and substituting it into the equation

$$(\hat{\mathbf{H}}^0+\hat{\mathbf{U}})\,\psi(\mathbf{r})=E\psi(\mathbf{r}), \tag{17.20}$$

and taking into account that $\psi^0_{\mathbf{k}}(\mathbf{r})$ and $\psi^0_{\mathbf{k}+2\pi\mathbf{g}}(\mathbf{r})$ are eigenfunctions of $\hat{\mathbf{H}}_0$ we obtain:

$$\alpha E^0(\mathbf{k})\,\psi^0_{\mathbf{k}}(\mathbf{r})+\beta E^0(\mathbf{k}+2\pi\mathbf{g})\,\psi^0_{\mathbf{k}+2\pi\mathbf{g}}(\mathbf{r})+U(\mathbf{r})\,\psi^0(\mathbf{r})=E\psi^0(\mathbf{r}). \tag{17.21}$$

Denote for the sake of simplicity

$$\begin{gathered}\psi^0_{\mathbf{k}}(\mathbf{r})=\psi_1;\quad \psi^0_{\mathbf{k}+2\pi\mathbf{g}}(\mathbf{r})=\psi_2;\\ E^0(\mathbf{k})=E_1;\quad E^0(\mathbf{k}+2\pi\mathbf{g})=E_2\end{gathered} \tag{17.22}$$

and rewrite (17.21) using the new notations:

$$\alpha E_1\psi_1+\beta E_2\psi_2+U(\alpha\psi_1+\beta\psi_2)=E(\alpha\psi_1+\beta\psi_2). \tag{17.23}$$

Premultiply these equations by ψ_1^* and ψ_2^* respectively, and integrate over $\mathbf{r}$ to obtain

$$\begin{gathered}\alpha E_1+\alpha U_{11}+\beta U_{12}=E\alpha,\\ \beta E_2+\alpha U_{21}+\beta U_{22}=E\beta,\end{gathered} \tag{17.24}$$

or

$$\begin{gathered}(E_1+U_{11}-E)\,\alpha+U_{12}\beta=0,\\ U_{21}\alpha+(E_2+U_{22}-E)\,\beta=0.\end{gathered} \tag{17.25}$$

The two equations (17.25) contain three unknowns E, α, β. *For the equations to have a non-trivial solution the determinant of the system should be zero.* This results in the third equation from which it is possible to calculate the energy E:

$$\begin{vmatrix}(E_1+U_{11}-E) & U_{12}\\ U_{21} & (E_2+U_{22}-E)\end{vmatrix}=0. \tag{17.26}$$

Solving this equation subject to the condition $U_{11}=U_{22}=\langle U\rangle=0$ we obtain

$$E=\frac{E_1+E_2}{2}\pm\sqrt{\frac{(E_1-E_2)^2}{4}+U_{12}U_{21}}. \tag{17.27}$$

In former notations this may be written as:

$$E^{(1)}(\mathbf{k})=\frac{E^0(\mathbf{k})+E^0(\mathbf{k}+2\pi\mathbf{g})}{2}\pm\sqrt{\frac{[E^0(\mathbf{k})-E^0(\mathbf{k}+2\pi\mathbf{g})]^2}{4}+|U_g|^2}. \tag{17.28}$$

We see from here that *when a perturbation* $W(\mathbf{r}) = U(\mathbf{r})$ *is applied the energy* $E^0(\mathbf{k})$ *displays a discontinuity at the point where* $E^0(\mathbf{k}) = E^0(\mathbf{k} + 2\pi\mathbf{g})$. *For these points* $E(\mathbf{k}) = E^0(\mathbf{k}) \pm |U_{\mathbf{g}}|$. *The value of the discontinuity is* $2|U_{\mathbf{g}}|$.

Let us investigate the states where the energy experiences discontinuity. The condition for discontinuity may be written in another form:

$$E^0(\mathbf{k}) - E^0(\mathbf{k} + 2\pi\mathbf{g}) = \frac{\hbar^2}{2m}[\mathbf{k}^2 - (\mathbf{k} + 2\pi\mathbf{g})^2] =$$

$$= \frac{\hbar^2}{2m}[4\pi^2\mathbf{g}^2 + 4\pi(\mathbf{k}\mathbf{g})] = 0. \qquad (17.29)$$

or

$$(\mathbf{k} + \pi\mathbf{g},\ \mathbf{g}) = 0. \qquad (17.30)$$

This condition has a simple geometrical and physical meaning. The problem is to find all values of $\mathbf{k}$ which satisfy the condition (17.30) This condition requires the vectors $\mathbf{g}$ and $\mathbf{k} + \pi\mathbf{g}$ to be orthogonal. Construct the vector $2\pi\mathbf{g}$. Place the origin of vector $\mathbf{k}$ at the point coinciding with the head of vector $2\pi\mathbf{g}$. Draw through the middle of the vector $2\pi\mathbf{g}$ a plane perpendicular to it (the plane BB' in Fig. 13). Any vector lying in this plane will be perpendicular to the vector $2\pi\mathbf{g}$, or $\mathbf{g}$. But a vector lying in the BB' plane may be obtained as a sum of the vectors $\pi\mathbf{g}$ and $\mathbf{k}$ if $\mathbf{k}$ terminates on the BB' plane, i.e., all the vectors $\mathbf{k}$ which terminate on the BB' plane satisfy the condition (17.30); this is the equation of a plane perpendicular to the vector $\mathbf{g}$. To find all possible values of $\mathbf{k}$ which satisfy the condition (17.19) all the vectors $\mathbf{g}$ should be specified. To this end one should construct a reciprocal lattice with the basic $2\pi\mathbf{b}_1$, $2\pi\mathbf{b}_2$, $2\pi\mathbf{b}_3$ superimposed on the $\mathbf{k}$-space, and draw vectors $2\pi\mathbf{g}$ connecting all the lattice sites with the origin of co-ordinates $\mathbf{k} = 0$. The family of planes passing through the middles of vectors $2\pi\mathbf{g}$ perpendicular to them defines all those vectors $\mathbf{k}$ which satisfy the condition

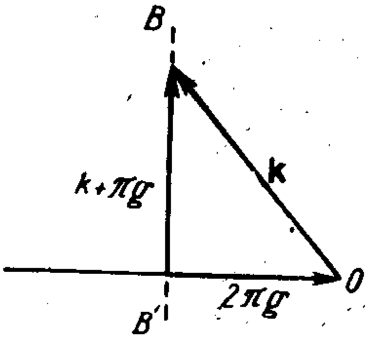


Fig. 13. The construction of a plane in which there is a discontinuity of energy

$$E^0(\mathbf{k} + 2\pi\mathbf{g}) = E^0(\mathbf{k}) \quad \text{or} \quad (\mathbf{k} + \pi\mathbf{g},\ \mathbf{g}) = 0. \qquad (17.31)$$

Thus, *equation* (17.30) *enables all Brillouin zones of any crystal to be constructed using its reciprocal lattice*. Figure 14 shows the method of constructing planes whose points satisfy equation (17.30) for various vectors $\mathbf{g}$. Different Brillouin zones have been striped differently. The construction of Brillouin zones on the basis of equation (17.30) is the most natural one.

To construct the main zone a minimum-volume polyhedron containing the point $\mathbf{k}=0$ should be isolated. Having found a polyhedron with next greatest volume and having cut out of it the first zone we will obtain the second Brillouin zone, etc.

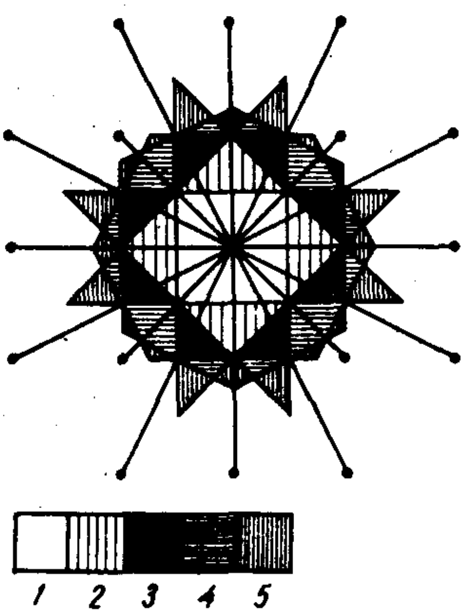


Fig. 14. The Brillouin zones of a plane square lattice

The reciprocal lattice for a simple cubic lattice is also simple cubic. If the origin $\mathbf{k}=0$ is placed in one of the lattice sites it will be surrounded by six nearest sites, they will give rise to six mutually orthogonal planes which will cut out of the wave-vector space a cube with the side $\frac{2\pi}{a}$ and the volume $\frac{8\pi^3}{a^3}$. The origin $\mathbf{k}=0$ in this case coincides with one of the lattice sites. The reciprocal lattice is constructed on the basis $2\pi\mathbf{b}_1$, $2\pi\mathbf{b}_2$, $2\pi\mathbf{b}_3$.

Consider the face-centered cubic lattice, its reciprocal lattice is volume-centered. If the origin $\mathbf{k}=0$ is placed in one of the cube apexes, the distance $\frac{2\pi}{a}$ along the axes will contain six sites arranged in the apexes of the cells. Six vectors issuing from these apexes determine the position of six mutually orthogonal planes, i.e. cut out a cube. The apexes of this cube are located at the lattice sites with the co-ordinates $(\pi\mathbf{b}, \pi\mathbf{b}, \pi\mathbf{b})$, $(-\pi\mathbf{b}, \pi\mathbf{b}, \pi\mathbf{b})$, etc.

If eight vectors are drawn from the apexes to the origin of co-ordinates they will determine eight planes perpendicular to the space diagonals of the cube. The planes cut off areas adjoining the apexes of the cube. The intersecting planes pass through points on the diagonals with the co-ordinates $\left(\frac{\pi}{a}, \frac{\pi}{a}, \frac{\pi}{a}\right)$, $\left(-\frac{\pi}{a}, \frac{\pi}{a}, \frac{\pi}{a}\right)$, etc. With the aid of these planes we arrive at a Brillouin zone shown in Fig. 15. This is a polyhedron with fourteen faces, six of which are squares, and eight—hexagons (cuboctahedron). In the same way Brillouin zones for other types of lattice may be constructed. The Brillouin zone for crystals with a diamond-type lattice coincides with the zone shown in Fig. 15. The first Brillouin zone for a volume-centered lattice is a dodecahedron.

Having discussed the geometrical meaning of equation (17.30) let us turn to its physical meaning. Taking into account that $|\mathbf{k}|=\frac{2\pi}{\lambda}$, where λ is the de Broglie wave length; $|\mathbf{b}|=\frac{l}{d}$, where l is an integer and d—the distance between neighbouring atomic planes of the direct lattice; denoting the angle between the vector $\mathbf{k}$ and the normal to the plane (the angle of incidence) by θ we may write the degeneracy condition (17.30) in the form:

$$(\mathbf{k}\mathbf{b}) + \pi\mathbf{b}^2 = 0 \tag{17.32}$$

or

$$-\frac{2\pi}{\lambda}\frac{l}{d}\cos\theta + \pi\frac{l^2}{d^2} = 0. \tag{17.33}$$

Cancelling out l, d and π we obtain

$$2d\cos\theta = l\lambda, \tag{17.34}$$

the well-known Wulf-Bragg condition for X-ray interference keeping in mind that θ is the angle of incidence and not Bragg (glancing) angle which usually stands in this expression.

Hence, *degeneracy will take place for states lying on the Brillouin zone boundaries — these are exactly the states for which Wulf-Bragg reflection should be observed.*

There will be two different combinations of wave functions $\psi_{\mathbf{k}}^0(\mathbf{r})$ and $\varphi_{\mathbf{k}+2\pi\mathbf{g}}^0(\mathbf{r})$ in the form of their sum and difference to correspond to two different energy values.

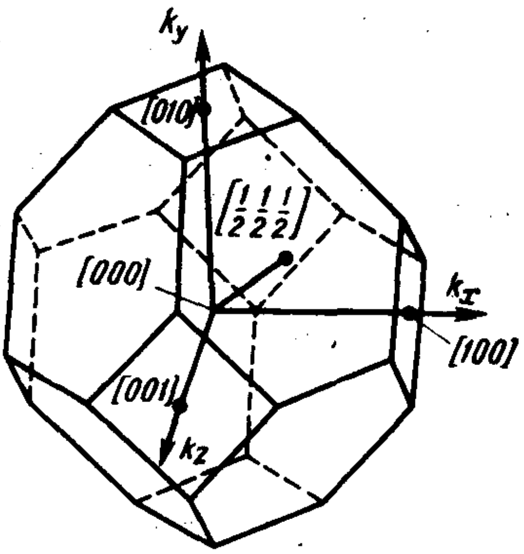


Fig. 15. The Brillouin zones of a face-centred cubic lattice. The co-ordinates of the points are given in units of $\frac{\pi}{a}$

The sum of two waves constitutes a single wave propagating in the direction of the vector $\pi\mathbf{g} + \mathbf{k}$, i.e. along the Brillouin zone boundary which is parallel to the respective atomic planes of the crystal. Thus, we arrive at a simple and visual interpretation of the causes of energy discontinuity. *Wulf-Bragg reflection results in the interference of the incident and reflected waves; standing waves arise in direction perpendicular to atomic planes; running waves are possible only along atomic planes.* Since some states are thereby forbidden so are also some energy values. Energy discontinuities arise on all Brillouin zone boundaries.

Perturbation is of significant importance also for the states near the Brillouin zone boundaries. Consider the dependence of energy on quasimomentum in the vicinity of discontinuity points where energy may be written in the form:

$$E(\mathbf{k}) = \frac{E^0(\mathbf{k}) + E^0(\mathbf{k} + 2\pi\mathbf{g})}{2} \pm \sqrt{\left[\frac{E^0(\mathbf{k}) - E^0(\mathbf{k} + 2\pi\mathbf{g})}{2}\right]^2 + |U_{\mathbf{g}}|^2}, \tag{17.35}$$

or

$$E(\mathbf{k})=\frac{\hbar^2\mathbf{k}^2}{2m}+\frac{\hbar^2\pi}{m}(\mathbf{g},\mathbf{k}+\pi\mathbf{g})\pm\sqrt{\left(\frac{\pi\hbar^2}{m}\right)^2(\mathbf{g},\mathbf{k}+\pi\mathbf{g})^2+|U_{\mathbf{g}}|^2}. \tag{17.36}$$

Denote the wave vector of points belonging to the Brillouin zone boundary by $\mathbf{k}_0$. Since $(\mathbf{g}, \mathbf{k}_0+\pi\mathbf{g})=0$,

$$E(\mathbf{k}_0)=\frac{\hbar^2\mathbf{k}_0^2}{2m}\pm|U_{\mathbf{g}}|. \tag{17.37}$$

This enables us to write for an arbitrary state $\mathbf{k}$:

$$E(\mathbf{k})-E(\mathbf{k}_0)=\frac{\hbar^2\mathbf{k}^2}{2m}-\frac{\hbar^2\mathbf{k}_0^2}{2m}+\frac{\hbar^2\pi}{m}(\mathbf{g},\mathbf{k}+\pi\mathbf{g})\pm$$
$$\pm\sqrt{\left(\frac{\pi\hbar^2}{m}\right)^2(\mathbf{g},\mathbf{k}+\pi\mathbf{g})^2+|U_{\mathbf{g}}|^2}\mp|U_{\mathbf{g}}|=f(\mathbf{k}). \tag{17.38}$$

Expand this expression into the Taylor series around the points $\mathbf{k}_0$. Obviously, $f(\mathbf{k}_0)=0$. Electron velocity

$$\mathbf{v}=\frac{dE(\mathbf{k})}{\hbar d\mathbf{k}}=\frac{\hbar\mathbf{k}}{m}+\frac{\pi\hbar\mathbf{g}}{m}\pm$$
$$\pm\frac{1}{\hbar}\left(\frac{\pi\hbar^2}{m}\right)^2\frac{(\mathbf{g},\mathbf{k}+\pi\mathbf{g})\,\mathbf{g}}{\sqrt{\left(\frac{\pi\hbar^2}{m}\right)^2(\mathbf{g},\underset{z}{\mathbf{k}}+\pi\mathbf{g})^2+|U_{\mathbf{g}}|^2}}. \tag{17.39}$$

The velocity of free motion (for states away from the Brillouin zone boundaries) is

$$\mathbf{v}=\frac{\hbar\mathbf{k}}{m}, \tag{17.40}$$

while in the vicinity of the boundaries the directions of velocity and of quasimomentum do not coincide. On the boundary itself $\mathbf{k}'=\mathbf{k}_0$. Therefore

$$\mathbf{v}=\frac{\hbar\mathbf{k}_0}{m}+\frac{\pi\hbar\mathbf{g}}{m}=\frac{\hbar}{m}(\mathbf{k}_0+\pi\mathbf{g}). \tag{17.41}$$

But the vector $\mathbf{k}_0+\pi\mathbf{g}$ lies in the same plane of the zone boundary. At the points of energy discontinuity velocity is directed along the boundary. Evidently, the predominant direction of velocity will lie in the boundary plane for the states in the near vicinity of the Brillouin zone boundary, as well. Since $\left.\frac{dE}{d\mathbf{k}}\right|_{\mathbf{k}_0}$ lies in the *boundary plane, surfaces of constant energy must be orthogonal to the Brillouin zone boundaries where they experience discontinuity*. For states $\mathbf{k}_0=-\pi\mathbf{g}$ velocity vanishes which means that motion in the direction perpendicular to the atomic planes is impossible. Since for these points $\frac{dE}{d\mathbf{k}}=0$, they must be the extremal points for energy. Find the second derivative

$$\frac{d^2E(\mathbf{k})}{\hbar^2 d\mathbf{k}^2}=\frac{1}{m}\pm\frac{\pi^2\hbar^2}{m^2}\frac{\mathbf{g}^2|U_{\mathbf{g}}|^2}{\left\{\left(\frac{\pi\hbar^2}{m}\right)^2(\mathbf{g},\mathbf{k}+\pi\mathbf{g})^2+|U_{\mathbf{g}}|^2\right\}^{\frac{3}{2}}}. \tag{17.42}$$

For points belonging to the Brillouin zone boundaries

$$\left.\frac{d^2E(\mathbf{k})}{\hbar^2 d\mathbf{k}^2}\right|_{\mathbf{k}_0}=\frac{1}{m}\left(1\pm\frac{\pi^2\hbar^2\mathbf{g}^2}{m|U_{\mathbf{g}}|}\right). \tag{17.43}$$

If the point $\mathbf{k}_0$ is an extremal point

$$\left.\frac{d^2E}{\hbar^2 d\mathbf{k}^2}\right|_{\mathbf{k}_0}=\frac{1}{m^*}=\frac{1}{m}\left(1\pm\frac{\pi^2\hbar^2\mathbf{g}^2}{m|U_{\mathbf{g}}|}\right), \tag{17.44}$$

or

$$m^*=\frac{m}{1\pm\frac{\pi^2\hbar^2\mathbf{g}^2}{m|U_{\mathbf{g}}|}}. \tag{17.45}$$

The plus sign corresponds to the energy $\frac{\hbar^2\mathbf{k}_0}{2m}+|U_{\mathbf{g}}|$, i.e. to the minimum value, the minus sign corresponds to the energy $\frac{\hbar^2\mathbf{k}_0}{2m}-|U_{\mathbf{g}}|$, i.e. to the maximum. Evaluate the order of magnitude of m^*. Setting $\hbar\approx 10^{-27}$ erg·s, $\pi g=10^8$ cm^{-1}, $m\approx 10^{-27}$ g we obtain

$$\frac{(\pi\hbar g)^2}{m}\approx 10^{-11}\ \text{erg}\approx 6\,\text{eV}. \tag{17.46}$$

If $|U_{\mathbf{g}}|$ is set at one electron-volt $m^*=\frac{m}{1\pm 6}$, i.e. for the energy minimum $m^*<0$, and for the maximum, $m^*>0$. At the same time $|m^*|$ is smaller than m, i.e. the effective mass is smaller than the free electron mass.

Now we will write out the expression for energy in the form of a Taylor series around an arbitrary point $\mathbf{k}_0$ of the Brillouin zone boundary:

$$E(\mathbf{k})=E(\mathbf{k}_0)+\frac{\hbar^2}{m}(\mathbf{k}_0+\pi\mathbf{g},\mathbf{k}-\mathbf{k}_0)+$$
$$+\frac{1}{2}\frac{\hbar^2}{m}\left(1\pm\frac{\pi^2\hbar^2\mathbf{g}^2}{m|U_{\mathbf{g}}|}\right)(\mathbf{k}-\mathbf{k}_0)^2+\ldots, \tag{17.47}$$

where

$$E(\mathbf{k}_0)=\frac{\hbar^2\mathbf{k}_0^2}{2m}\pm|U_{\mathbf{g}}|.$$

The expansion starts with a linear term. If the vector $\mathbf{k}-\mathbf{k}_0$ is perpendicular to the zone boundary energy will be a quadratic function of $(\mathbf{k}-\mathbf{k}_0)$.

Now let us discuss the main problem—the formation of energy bands. Here we must take into account the fact that we obtained two branches of energy for the points lying near the Brillouin zone boundary:

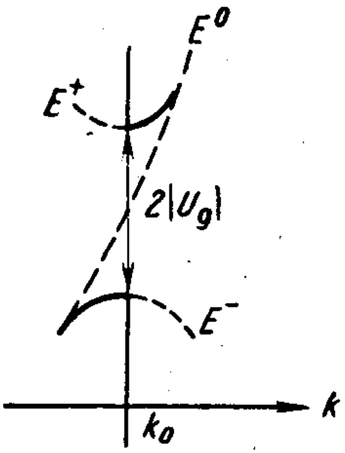


Fig. 16. Two branches of energy E^+ and E^- in the vicinity of the Brillouin zone

$$E^+(\mathbf{k}) = \frac{\hbar^2\mathbf{k}_0^2}{2m} + |U_g| + \frac{\hbar^2}{m}(\mathbf{k}_0+\pi\mathbf{g}, \mathbf{k}-\mathbf{k}_0) + \frac{\hbar^2}{2m}\left(1+\frac{\pi^2\hbar^2\mathbf{g}^2}{m|U_g|}\right)(\mathbf{k}-\mathbf{k}_0)^2,$$

$$E^-(\mathbf{k}) = \frac{\hbar^2\mathbf{k}_0^2}{2m} - |U_g| + \frac{\hbar^2}{m}(\mathbf{k}_0+\pi\mathbf{g}, \mathbf{k}-\mathbf{k}_0) + \frac{\hbar^2}{2m}\left(1-\frac{\pi^2\hbar^2\mathbf{g}^2}{m|U_g|}\right)(\mathbf{k}-\mathbf{k}_0)^2. \qquad (17.48)$$

Both functions E^+ and E^- are continuous on the Brillouin zone boundary. Figure 16 shows the dependence of E^+ and E^- on $(\mathbf{k}-\mathbf{k}_0)$ for directions along the normal to the zone boundary. For energy to experience discontinuity at the boundary we should presume each expression E^+ and E^- to be valid for one side of the boundary only. Since inside each zone energy is continuous and since E^+ and E^- away from the boundary must assume the form $E^0(\mathbf{k}) = \frac{\hbar^2\mathbf{k}^2}{2m}$, we must choose E^- for the *internal side of the boundary* and E^+ for the *external side of the boundary* (internal side is the one which contains the origin $\mathbf{k}=0$), the branches E^+ and E^- are shown on Fig. 16 by solid lines.

Energy discontinuity at points $\mathbf{k}_0$ is not tantamount to discontinuity in the energy spectrum $E(\mathbf{k})$. To prove it take two points $\mathbf{k}_1$ and $\mathbf{k}_2$ very close to the internal and external sides of the boundary, respectively, such that when $|\mathbf{k}_1| > |\mathbf{k}_2|$

$$\frac{\hbar^2\mathbf{k}_1^2}{2m} - |U_g| > \frac{\hbar^2\mathbf{k}_2^2}{2m} + |U_g|, \qquad (17.49)$$

i.e. the energy bands E^- and E^+ overlap. But this means that in this case the energy spectrum is continuous.

For a forbidden band to arise in the energy spectrum there must be no overlapping to the bands. This possibility becomes obvious if other boundary surfaces are taken into the picture.

Consider qualitatively the nature of the energy spectrum near one of the Brillouin zone apexes. Suppose zone boundaries are defined by three equations.

$$\left.\begin{aligned}(\mathbf{g}_1,\ \mathbf{k}_{01}+\pi\mathbf{g}_1)=0,\\(\mathbf{g}_2,\ \mathbf{k}_{02}+\pi\mathbf{g}_2)=0,\\(\mathbf{g}_3,\ \mathbf{k}_{03}+\pi\mathbf{g}_3)=0,\end{aligned}\right\}\qquad(17.50)$$

where $\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$ are vectors of the reciprocal lattice, $\mathbf{k}_{01}$, $\mathbf{k}_{02}$, $\mathbf{k}_{03}$, current radius vectors of respective boundary planes.

Since the velocities corresponding to states belonging to boundary planes lie in these planes, the velocity for states arranged along the line of intersection of two planes may be directed only along this line.

Hence, the velocity $\mathbf{v}(\mathbf{k}_0^0)$ for a state corresponding to the point of intersection $\mathbf{k}_0^0$ of three planes should be zero, i.e. there should be an energy extremum in the apex of the Brillouin zone, since at that point

$$\left.\frac{dE}{d\mathbf{k}}\right|_{\mathbf{k}_0^0}=0\qquad(17.51)$$

Consequently, the Taylor series for extrema lying at zone apexes will not contain linear terms. Assuming the vectors $\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$ to be mutually orthogonal we can derive an approximate expression for energy in the vicinity of point $\mathbf{k}_0^0$. Three equations should be satisfied simultaneously:

$$E^{\pm}(\mathbf{k})=E^{\pm}(\mathbf{k}_0^0)+\frac{1}{2}\frac{\hbar^2}{m}\left(1\pm\frac{\pi^2\hbar^2g_i^2}{m\,|U_{g_i}|}\right)(\mathbf{k}-\mathbf{k}_0^0)^2\ \ (i=1,\ 2,\ 3).\qquad(17.52)$$

This will be the case only if each of them is satisfied for projections of $(\mathbf{k}-\mathbf{k}_0^0)$ on corresponding $\mathbf{g}_i$ vectors so that

$$E^{\pm}(\mathbf{k})=E^{\pm}(\mathbf{k}_0^0)\pm\sum_{i=1}^{3}\frac{\hbar^2}{2m}\left(1\pm\frac{\pi^2\hbar^2g_i^2}{m\,|U_{g_i}|}\right)(k_i-k_{0i}^0)^2=$$

$$=E^{\pm}(\mathbf{k}_0^0)\pm\sum_{i=1}^{3}\frac{\hbar^2}{2m_i^*}(k_i-k_{0i}^0)^2.\qquad(17.53)$$

Here plus and minus should relate to states internal and external with respect to each surface. The value of discontinuity will be different for different directions. Since thermal forbidden band is determined by the minimum energy gap, essential to electric conductivity phenomena,, it will be determined by the minimum of the three $(2|U_{g_i}|)$. For phenomena for which the quasimomentum conservation law is essential the forbidden band will be different and for this reason the optical forbidden band should vary from the minimum to the maximum of the $(2|U_{g_i}|)$

The fact that there is actually a discontinuity in the energy spectrum becomes evident from the expression for energy itself, away from the point $\mathbf{k}_0^0$ the energy increases as compared to $E^+(\mathbf{k}_0^0)$, or decreases as compared to $E^-(\mathbf{k}_0^0)$, $E(\mathbf{k}_0^0)$ not being able to change in this case. At the same time, when energy becomes discontinuous at points in boundary planes, its changes are due to the changes of $E(\mathbf{k}_0)$ and of the linear term.

To obtain a more accurate expression for energy in the vicinity of Brillouin zone apexes quadruple energy degeneration should be taken into account, and to this end zero-approximation wave function should be taken in the form

$$\psi^0(\mathbf{r}) = \sum_{i=1}^{4} \alpha_i \psi^0_{\mathbf{k}_i}(\mathbf{r}), \tag{17.54}$$

where $\psi^0_{\mathbf{k}_i}(\mathbf{r})$ is of the form

$$\psi^0_{\mathbf{k}_1}(\mathbf{r}) = \frac{1}{\sqrt{G}} e^{i(\mathbf{kr})};$$

$$\psi^0_{\mathbf{k}_j}(\mathbf{r}) = \frac{1}{\sqrt{G}} e^{i(\mathbf{k}+2\pi\mathbf{g}_j;\,\mathbf{r})} \quad (j = 1,\ 2,\ 3). \tag{17.55}$$

We have obtained a fourth-order equation for $E(\mathbf{k})$.

Finally we will discuss the shape of constant-energy surfaces in the Brillouin zones and the dependence of energy on $|\mathbf{k}|$ along certain directions.

Constant-energy surfaces near the centre of the first Brillouin zone for $\mathbf{k}=0$ are spheres. Figure 17 shows energy surfaces, i.e. constant-energy lines, for a plane two-dimensional lattice. As the distance from the centre of the spheres increases their deformation sets in. The nature of the deformation may be understood if one considers energy surfaces in zone corners where they are part of ellipsoid surface.

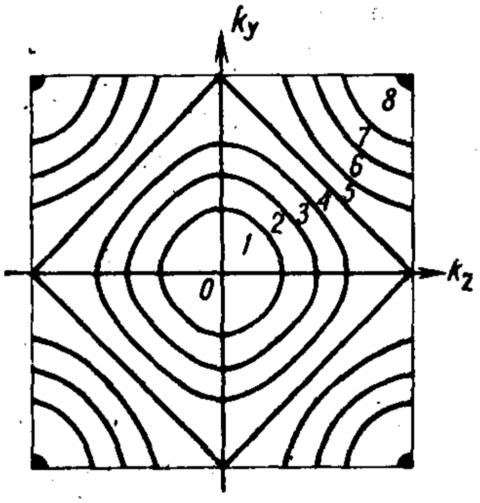


Fig. 17. Constant-energy lines for a square plane lattice

Consider now the dependence of energy on $|\mathbf{k}|$ for some direction. Since constant-energy surfaces are of an intricate shape the $E(|\mathbf{k}|)$ graphs will be different for different directions. Figure 18 shows $E(\mathbf{k})$ dependence for one of the axial directions perpendicular to the Brillouin zone boundaries. Dotted line shows the $E^0(\mathbf{k})$ dependence for the free electron. For different directions the distance to the boundary is different, and for this reason the position of the point where energy becomes discontinuous depends on direction and is of varying magnitude. The energy dependence may be extrapolated into neighbouring zones as is shown by dotted line in Fig. 18, and this brings home the already known fact that energy is a periodic function of **k**:

$$E(\mathbf{k}+2\pi\mathbf{b})=E(\mathbf{k}). \tag{17.56}$$

A different representation is, however, possible. Because of periodic dependence of energy on the wave vector the parts of the curves shown in Fig. 18 by solid lines may be transferred into the first Brillouin zone. In this case energy becomes a multivalued function of the wave vector. However, for each band energy is single-valued. (In real crystals energy may be multivalued even for one band.) These graphs may conveniently be re-drawn so as to place the extrema of two neighbouring zones into the centre of the first Brillouin zone (Fig. 19). As we will see below *in real crystals energy extrema for different zones may be at different sites of the Brillouin zone.*

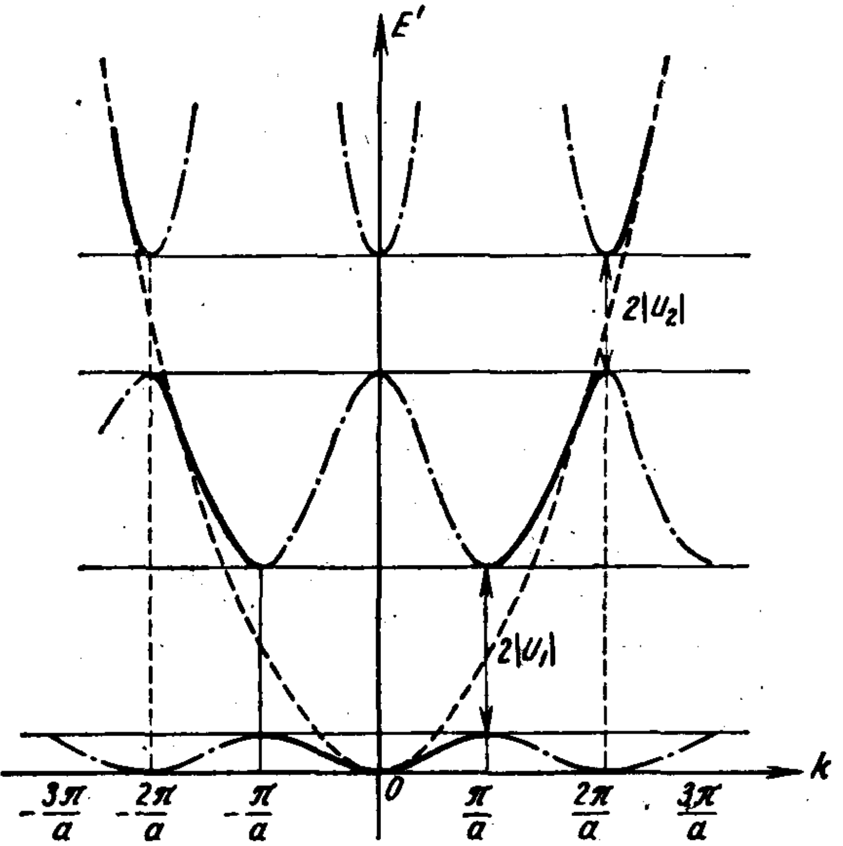


Fig. 18. The dependence of energy on the wave vector for a fixed crystallographic direction

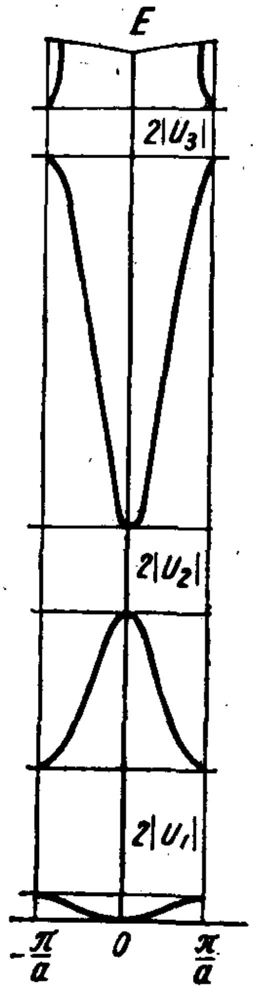


Fig. 19. The energy spectrum in the main Brillouin zone

Summary of Sec. 17

1. The theory of the quasifree electron accepts the solution of the Schrödinger equation for the free electron as the zero approximation for the problem of the electron in a periodic field:

$$\hat{H}^0 = \hat{T} = -\frac{\hbar^2}{2m}\Delta; \quad E^0(\mathbf{k}) = \frac{\hbar^2 \mathbf{k}^2}{2m}; \tag{17.1s}$$

$$\psi_{\mathbf{k}}^0(\mathbf{r}) = \frac{1}{\sqrt{G}} e^{i(\mathbf{kr})}. \tag{17.2s}$$

2. The potential energy of the lattice field is regarded as a perturbation

$$\hat{W}(\mathbf{r}) = \hat{U}(\mathbf{r}). \tag{17.3s}$$

The matrix elements of the operator **W** calculated with the aid of wave functions $\psi_{\mathbf{k}}^0(\mathbf{r})$ are not equal to zero when $\mathbf{k}' = \mathbf{k} + 2\pi\mathbf{b}$, in this case they are equal to the Fourier series coefficients of $U(\mathbf{r})$ expansion

$$W_{\mathbf{k}'\mathbf{k}} = U_{\mathbf{b}} = c_{\mathbf{b}}; \quad \mathbf{b} = \frac{\mathbf{k}' - \mathbf{k}}{2\pi}. \tag{17.4s}$$

For small values of $|\mathbf{b}|$ the value of $U_{\mathbf{b}}$ will be greater than for large $|\mathbf{b}|$ since large values of $|\mathbf{b}|$ correspond to higher "harmonics" of the field $U(\mathbf{r})$.

3. The first approximation results in the downward displacement of the energy spectrum by $\langle U \rangle$. Energy correction for most states in the second approximation is practically zero. However, there are such states for which the corrections to energy and wave functions become infinite. Physically, this means that such states are degenerate.

4. Degenerate **k** states lie in planes that satisfy the equation

$$(\mathbf{b},\ \mathbf{k} + \pi\mathbf{b}) = 0 \tag{17.5s}$$

which is the Wulf-Bragg equation. These planes are conveniently chosen as Brillouin zone boundaries. The solution of the problem by the theory of perturbations method yields an energy discontinuity of $2|U_{\mathbf{b}}|$ at the degeneracy points, i.e. on the Brillouin zone boundaries.

5. Energy discontinuity results in the discontinuity of the energy spectrum. Forbidden energy values are termed forbidden bands. The forbidden band width is equal to $2|U_{\mathbf{b}}|$, hence it increases with the increase of $U(\mathbf{r})$. As energy E increases the forbidden band width decreases, but the energy band width increases (Fig. 19).

6. The physical cause of the band structure of the energy spectrum is the Wulf-Bragg interference of electron waves.

7. The effective mass is of different signs at the maximum and the minimum of energy. Neglecting unity as compared to other terms in the denominator of expression (17.45) we obtain

$$\frac{m^*}{m} = \pm \frac{m}{\pi^2\hbar^2 g^2} \left| U_g \right|. \tag{17.6s}$$

It follows from (17.6s) that: (a) $|m^*|$ is proportional to the forbidden band width, (b) as energy increases $|m^*|$ decreases, i.e. in the higher energy bands $|m^*|$ is smaller than in the lower ones.

8. In the vicinity of extrema the energy is a quadratic function of the wave vector. Energy surfaces are orthogonal to the Brillouin zone boundaries.

18. THEORY OF THE QUASIBOUND ELECTRON

The theory of the quasibound electron accepts electron state of an isolated atom as zero approximation for the solution of the Schrödinger equation for the electron in a periodic field,

$$\left. \begin{aligned} \hat{\mathbf{H}}\psi(\mathbf{r}) &= E\psi(\mathbf{r}), \\ \hat{\mathbf{H}} = -\frac{\hbar^2}{2m}\Delta + \mathbf{U}(\mathbf{r}); &\quad \mathbf{U}(\mathbf{r}) = \mathbf{U}(\mathbf{r}+\mathbf{n}). \end{aligned} \right\} \tag{18.1}$$

Denote the Hamiltonian of an isolated atom by

$$\hat{\mathbf{H}}_a = -\frac{\hbar^2}{2m}\Delta + V_a(\mathbf{r}) \tag{18.2}$$

and write the Schrödinger equation for it

$$\hat{\mathbf{H}}_a\psi_a(\mathbf{r}) = E_a\psi_a(\mathbf{r}), \tag{18.3}$$

where $V_a(\mathbf{r})$ is the potential energy of the electron in an isolated atom, E_a—some energy level, $\psi_a(\mathbf{r})$—wave function corresponding to E_a. Figure 20 shows a diagram of energy levels of an isolated atom. An essential property of the atomic wave function is its sharp dependence on the distance: as r exceeds a certain value it decreases exponentially. The solution of the equation for the atom is supposed to be known to the reader. To find the first approximation for the electron energy in a crystal a zero approximation for the wave function should be chosen. Suppose at first that the atoms in the crystal do not interact. Place the origin of co-ordinates in one of the atom sites. Then it will be possible to define the co-ordinates of other atoms in the form of a translation vector:

$$\mathbf{n} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3.$$

If the current radius vector is denoted by $\mathbf{r}$ the distance between the given point $\mathbf{r}$ and the ion $\mathbf{n}$ will be equal to $|\mathbf{r}-\mathbf{n}|$, and the electron wave function in the nth atom will assume the form $\psi_a(\mathbf{r}-\mathbf{n})$.

The electron wave function in a crystal $\psi^0(\mathbf{r})$ should be the sum of atomic wave functions $\psi_a(\mathbf{r}-\mathbf{n})$:

$$\psi^0(\mathbf{r})=\sum_{\mathbf{m}} c_{\mathbf{m}}\psi_a(\mathbf{r}-\mathbf{m}). \tag{18.4}$$

The coefficients $c_{\mathbf{m}}$ are chosen to make $\psi^0(\mathbf{r})$ satisfy the translational condition

$$\psi^0(\mathbf{r}+\mathbf{n})=e^{i(\mathbf{kn})}\psi^0(\mathbf{r}), \tag{18.5}$$

where $\mathbf{n}$ is the translation vector, and $\mathbf{k}$—the wave vector. For the condition of translation to be satisfied the coefficients $c_{\mathbf{m}}$ may be chosen in the form

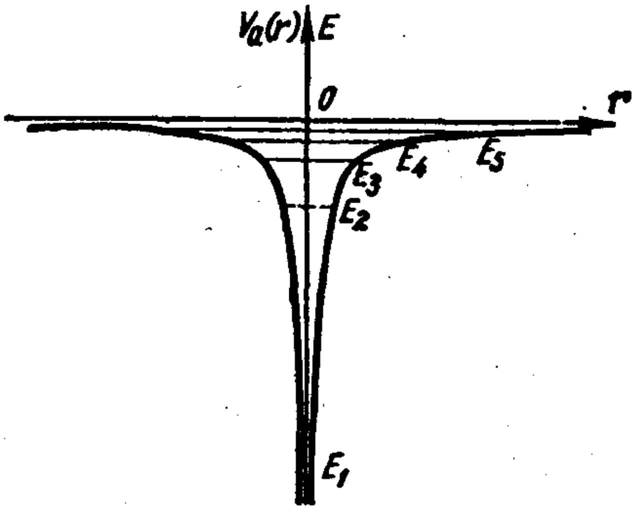


Fig. 20. The potential energy and energy levels of an isolated atom

$$c_{\mathbf{m}}=e^{i(\mathbf{km})}, \tag{18.6}$$

or

$$\psi^0(\mathbf{r})=\sum_{\mathbf{m}} e^{i(\mathbf{km})}\psi_a(\mathbf{r}-\mathbf{m}). \tag{18.7}$$

Check whether the condition (18.5) is satisfied for the function (18.7):

$$\psi^0(\mathbf{r}+\mathbf{n})=\sum_{\mathbf{m}} e^{i(\mathbf{km})}\psi_a(\mathbf{r}-\mathbf{m}+\mathbf{n})=$$
$$=e^{i(\mathbf{kn})}\sum_{\mathbf{m}} e^{i(\mathbf{k},\,\mathbf{m}-\mathbf{n})}\psi_a[\mathbf{r}-(\mathbf{m}-\mathbf{n})]=e^{i(\mathbf{kn})}\sum_{\mathbf{l}} e^{i(\mathbf{kl})}\psi_a(\mathbf{r}-\mathbf{l}). \tag{18.8}$$

Evidently, $\mathbf{l}=\mathbf{m}-\mathbf{n}$ assumes the same values as $\mathbf{m}$ but in a different sequence, therefore

$$\sum_{\mathbf{l}} e^{i(\mathbf{kl})}\psi_a(\mathbf{r}-\mathbf{l})=\psi^0(\mathbf{r}), \tag{18.9}$$

and the condition of translation is thus satisfied.

Since $\psi_a(\mathbf{r}-\mathbf{l})$ are normalized to unity, $\psi^0(\mathbf{r})$ is not normalized. Find its normalizing factor:

$$\int \psi^{0*}(\mathbf{r})\,\psi^0(\mathbf{r})\,d\tau = \int \left[\sum_{\mathbf{m}} e^{-i(\mathbf{km})}\psi_a^*(\mathbf{r}-\mathbf{m})\right]\left[\sum_{\mathbf{n}} e^{i(\mathbf{kn})}\psi_a(\mathbf{r}-\mathbf{n})\right] d\tau =$$
$$=\sum_{\mathbf{m}}\sum_{\mathbf{n}} e^{i(\mathbf{k},\,\mathbf{n}-\mathbf{m})} \int \psi_a^*(\mathbf{r}-\mathbf{m})\,\psi_a(\mathbf{r}-\mathbf{n})\,d\tau. \qquad (18.10)$$

The double sum is easily reduced to a single sum. To this end denote

$$\mathbf{r}-\mathbf{m}=\mathbf{r}'; \quad \mathbf{r}=\mathbf{r}'+\mathbf{m}; \quad d\tau=d\tau', \qquad (18.11)$$

then

$$\int \psi_a^*(\mathbf{r}-\mathbf{m})\,\psi_a(\mathbf{r}-\mathbf{n})\,d\tau = \int \psi_a^*(\mathbf{r}')\,\psi_a[\mathbf{r}'-(\mathbf{n}-\mathbf{m})]\,d\tau'. \qquad (18.12)$$

The integral of the product of wave functions $\psi_a^*(\mathbf{r}')$ and $\psi_a[\mathbf{r}'-(\mathbf{n}-\mathbf{m})]$ depends not on the position of the ions $\mathbf{n}$ and $\mathbf{m}$ but only on the distance $\mathbf{t}=\mathbf{n}-\mathbf{m}$ between them. Denote

$$\int \psi_a^*(\mathbf{r}')\,\psi_a(\mathbf{r}'-\mathbf{t})\,d\tau' = S_{\mathbf{t}}. \qquad (18.13)$$

When $\mathbf{t}=0$, $S_{\mathbf{t}}=1$. If the distance between the ions is large the wave functions will not overlap, and $S_{\mathbf{t}}$ will be practically zero. $S_{\mathbf{t}}$ are non-zero only for small $\mathbf{t}$'s.

Since for fixed $\mathbf{m}$ $\mathbf{n}-\mathbf{m}$ and $\mathbf{t}$ assume the same values the sum over $\mathbf{n}$ will be equal to the sum over $\mathbf{t}$:

$$\sum_{\mathbf{n}} e^{i(\mathbf{k},\,\mathbf{n}-\mathbf{m})} \int \psi_a^*(\mathbf{r}-\mathbf{m})\,\psi_a(\mathbf{r}-\mathbf{n})\,d\tau = \sum_{\mathbf{t}} e^{i(\mathbf{kt})} S_{\mathbf{t}}. \qquad (18.14)$$

The second summation sign embraces similar terms of the form $\left[\sum_{\mathbf{t}} e^{i(\mathbf{kt})} S_{\mathbf{t}}\right]$, their number being equal to N—the number of atoms in the crystal. Therefore

$$\int \psi^{*0}(\mathbf{r})\,\psi^0(\mathbf{r})\,d\tau = N \sum_{\mathbf{t}} e^{i(\mathbf{kt})} S_{\mathbf{t}}. \qquad (18.15)$$

Find the form of the perturbation operator.

Suppose the crystal is "elongated" so that atoms cease to interact. In this case the *periodic potential field of the crystal is made up of the potential fields of isolated atoms periodically repeated* (Fig. 21), and therefore

$$U(\mathbf{r}) = \sum_{\mathbf{n}} V_a(\mathbf{r}-\mathbf{n}). \qquad (18.16)$$

It is peculiar to this sum that at each point $\mathbf{r}$ the function $U(\mathbf{r})$ is determined only by the shape of the potential curve of the nearest atom. When the atoms of the crystal are brought together the sum of potential curves of isolated atoms at every point **r** will, as before, make up the potential energy of the electron in the combined field of all atoms (Fig. 22). However, in such a case atomic interaction will be neglected, or, in other words, the sum will not be self-consistent. Suppose self-consistency of lattice potential field is achieved by adding some energy $W(\mathbf{r})$ in such a way that

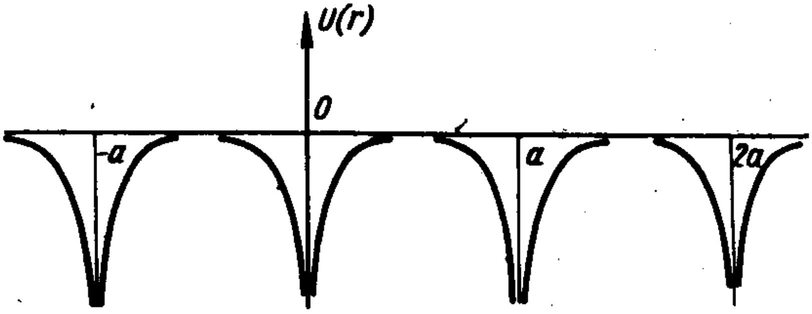


Fig. 21. The lattice field of an "extended" crystal

$$U(\mathbf{r})=\sum_{\mathbf{n}} V_a(\mathbf{r}-\mathbf{n})+W(\mathbf{r}). \tag{18.17}$$

The function $W(\mathbf{r})$ is periodic since both $U(\mathbf{r})$ and $\sum_{\mathbf{n}} V_a(\mathbf{r}-\mathbf{n})$ are periodic functions, $W(\mathbf{r})$ has no singular points, the singular points of $U(\mathbf{r})$ are determined by the atomic potential curves $V_a(\mathbf{r}-\mathbf{n})$. In other words, $W(\mathbf{n})$ is close to zero, and $W(\mathbf{r})$ is limited everywhere. The analytical expression for $W(\mathbf{r})$ can be found quite easily:

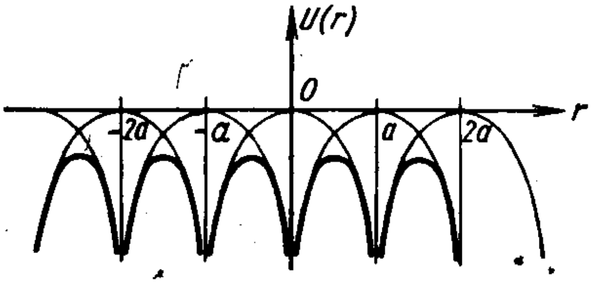


Fig. 22. The crystal lattice field with no account taken of the electron shell interaction

$$W(\mathbf{r})=U(\mathbf{r})-\sum_{\mathbf{n}} V_a(\mathbf{r}-\mathbf{n}) \tag{18.18}$$

with the aid of "known" $U(\mathbf{r})$ and $V_a(\mathbf{r}-\mathbf{n})$. As long as atoms are attracted $W(\mathbf{r})$ will be negative. When as a result of a very strong contraction of the crystal the atoms will begin to be repelled, $W(\mathbf{r})$ will turn positive.

Now we will have to find the energy. As is well known, the zero approximation of the wave function ψ^0 enables the first approximation for energy $E^{(1)}$ to be obtained:

$$\left[-\frac{\hbar^2}{2m}\Delta+\sum_{\mathbf{n}} V_a(\mathbf{r}-\mathbf{n})+W(\mathbf{r})\right]\psi^0(\mathbf{r})=E^{(1)}\psi^0(\mathbf{r}). \qquad (18.19)$$

Multiplying the equation (18.19) by $\psi^{0*}(\mathbf{r})$ and integrating over the crystal volume we obtain

$$E^{(1)}=\frac{\int \psi^{0*}(\mathbf{r})\left[-\frac{\hbar^2}{2m}\Delta+\sum_{\mathbf{n}} V_a(\mathbf{r}-\mathbf{n})+W(\mathbf{r})\right]\psi^0(\mathbf{r})\,d\tau}{\int \psi^{0*}(\mathbf{r})\,\psi^0(\mathbf{r})\,d\tau}. \qquad (18.20)$$

The denominator of the expression (18.20) was determined previously [see (18.15)]. Calculate the numerator which we will denote by **I**:

$$\begin{aligned}\mathbf{I}&=\int \psi^{0*}(\mathbf{r})\left[-\frac{\hbar^2}{2m}\Delta+\sum_{\mathbf{n}} V_a(\mathbf{r}-\mathbf{n})+W(\mathbf{r})\right]\psi^0(\mathbf{r})\,d\tau=\\&=\int\left\{\left[\sum_{\mathbf{l}} e^{-i(\mathbf{kl})}\psi_a^*(\mathbf{r}-\mathbf{l})\right]\left[-\frac{\hbar^2}{2m}\Delta+\sum_{\mathbf{n}} V_a(\mathbf{r}-\mathbf{n})+W(\mathbf{r})\right]\times\right.\\&\quad\times\left.\left[\sum_{\mathbf{t}} e^{i(\mathbf{kt})}\psi_a(\mathbf{r}-\mathbf{t})\right]\right\}d\tau=\left\{\sum_{\mathbf{l},\mathbf{t}} e^{i\mathbf{k}(\mathbf{t}-\mathbf{l})}\int\psi_a^*(\mathbf{r}-\mathbf{l})\times\right.\\&\quad\times\left.\left[-\frac{\hbar^2}{2m}\Delta+\sum_{\mathbf{n}} V_a(\mathbf{r}-\mathbf{n})+W(\mathbf{r})\right]\psi_a(\mathbf{r}-\mathbf{t})\right\}d\tau, \qquad (18.21)\end{aligned}$$

Put

$$\begin{aligned}&\mathbf{t}-\mathbf{l}=\mathbf{p};\ \mathbf{r}=\mathbf{r}'+\mathbf{l};\ \mathbf{r}-\mathbf{l}=\mathbf{r}';\ d\tau=d\tau';\\&\mathbf{n}'=\mathbf{n}-\mathbf{l};\ W(\mathbf{r})=W(\mathbf{r}').\end{aligned} \qquad (18.22)$$

In this case the integral **I** may be written in the form of single sums

$$\begin{aligned}\mathbf{I}&=N\left\{\sum_{\mathbf{p}} e^{i(\mathbf{kp})}\int\psi_a^*(\mathbf{r}')\times\right.\\&\times\left.\left[-\frac{\hbar^2}{2m}\Delta'+\sum_{\mathbf{n}'} V_a(\mathbf{r}'-\mathbf{n}')+W(\mathbf{r}')\right]\psi_a(\mathbf{r}'-\mathbf{p})\,d\tau'\right\} \qquad (18.23)\end{aligned}$$

as it was done when the normalizing term for $\psi^0(\mathbf{r})$ was calculated.

Taking into account that

$$\left[-\frac{\hbar^2}{2m}\Delta'+V_a(\mathbf{r}'-\mathbf{p})+\left\{\sum_{\mathbf{n}'\neq\mathbf{p}}V_a(\mathbf{r}'-\mathbf{n}')+W(\mathbf{r}')\right\}\right]\psi_a(\mathbf{r}'-\mathbf{p})=$$
$$=E_a\psi_a(\mathbf{r}'-\mathbf{p})+\left\{\sum_{\mathbf{n}'\neq\mathbf{p}}V_a(\mathbf{r}'-\mathbf{n}')+W(\mathbf{r}')\right\}\psi_a(\mathbf{r}'-\mathbf{p}) \quad (18.24)$$

we may write

$$\mathbf{I}=N\sum_{\mathbf{p}}e^{i(\mathbf{kp})}\int\psi_a^*(\mathbf{r}')\times$$
$$\times\left[E_a+\left\{\sum_{\mathbf{n}'\neq\mathbf{p}}V_a(\mathbf{r}'-\mathbf{n}')+W(\mathbf{r}'-\mathbf{p})\right\}\right]\psi_a(\mathbf{r}'-\mathbf{p})\,d\tau'=$$
$$=NE_a\sum_{\mathbf{p}}e^{i(\mathbf{kp})}S_{\mathbf{p}}+N\sum_{\mathbf{p}}e^{i(\mathbf{kp})}\int\psi_a^*(\mathbf{r}')\times$$
$$\times\left\{\sum_{\mathbf{n}'\neq\mathbf{p}}V_a(\mathbf{r}'-\mathbf{n}')+W(\mathbf{r}')\right\}\psi_a(\mathbf{r}'-\mathbf{p})\,d\tau'. \quad (18.25)$$

Take out of the second addend the term with $\mathbf{p}=0$ and denote it by C

$$C=\int\psi_a^*(\mathbf{r}')\left\{\sum_{\mathbf{n}'\neq 0}V_a(\mathbf{r}'-\mathbf{n}')+W(\mathbf{r}')\right\}\psi_a(\mathbf{r}')\,d\tau'. \quad (18.26)$$

The quantity C is the averaged potential energy of the electron belonging to an atom in the field of all other atoms which takes into account the self-consistency of the field. Now the expression for $\mathbf{I}$ may be written in the form:

$$\mathbf{I}=NE_a\sum_{\mathbf{p}}e^{i(\mathbf{kp})}S_{\mathbf{p}}+NC+$$
$$+N\sum_{\mathbf{p}}e^{i(\mathbf{kp})}\int\psi_a^*(\mathbf{r}')\left[\sum_{\mathbf{n}'\neq\mathbf{p}}V_a(\mathbf{r}'-\mathbf{n}')+W(\mathbf{r}')\right]\psi_a(\mathbf{r}'-\mathbf{p})\,d\tau'. \quad (18.27)$$

Denote the integral in (18.27) by $A(\mathbf{p})$:

$$\int\psi_a^*(\mathbf{r}')\left[\sum_{\mathbf{n}'\neq\mathbf{p}}V_a(\mathbf{r}'-\mathbf{n}')+W(\mathbf{r}')\right]\psi_a(\mathbf{r}'-\mathbf{p})\,d\tau'=A(\mathbf{p}). \quad (18.28)$$

$A(\mathbf{p})$ is the exchange energy the origin of which is the fact that the electron may, with some probability, be located near any atom. Evidence for this is that $A(\mathbf{p})$ is made up of the wave functions of two atoms separated by the distance $|\mathbf{p}|$. In other words, two atoms at a distance $|\mathbf{p}|$ from each other may exchange electrons. The exchange proceeds via the field of all other atoms and the periodic self-consistent part of the lattice field $W(\mathbf{r})$. Obviously,

$A(\mathbf{p})$ is not negligible only for small $|\mathbf{p}|$'s since because of exponential nature of the wave functions for large $|\mathbf{p}|$'s $A(\mathbf{p}) \approx 0$. Physically, this means that *the exchange takes place mainly between nearest atoms. Electron exchange between any two atoms of the crystal takes place by way of a chain of nearest neighbour exchanges.* In other words, *electrons are not localized near individual atoms but move "freely" through the crystal jumping from one atom to another by the exchange process.* It should be noted, besides, that the main part in $A(\mathbf{p})$ should be played by the periodic field $W(\mathbf{r})$ which reaches a considerable value in the vicinity of every atom, while $V_a(\mathbf{r}'-\mathbf{n}')$ in the vicinity of atoms taking part in the "exchange" is small ,for $\mathbf{n}' \neq 0$ and $\mathbf{n}' \neq \mathbf{p}$.

With the expression for I in mind we may write the expression for $E^{(1)}$:

$$E^{(1)'} = \frac{N\left[E_a \sum_{\mathbf{p}} e^{i(\mathbf{kp})} S_{\mathbf{p}} + C + \sum_{\mathbf{p}} e^{i(\mathbf{kp})} A(\mathbf{p})\right]}{N \sum_{\mathbf{p}} e^{i(\mathbf{kp})} S_{\mathbf{p}}} =$$

$$= E_a + C\left[\sum_{\mathbf{p}} e^{i(\mathbf{kp})} S_{\mathbf{p}}\right]^{-1} + \frac{\sum_{\mathbf{p}} e^{i(\mathbf{kp})} A(\mathbf{p})}{\sum_{\mathbf{p}} e^{i(\mathbf{kp})} S_{\mathbf{p}}}. \qquad (18.29)$$

We see from here that *the energy of the electrons of an atomic system is, in the zero approximation, equal to the energy of electrons of separate atoms.* As the atoms are brought together (i.e. when the interaction of the electrons with the ions and with other electrons is taken into account) *energy levels are lowered by the amount C and split into a band of a definite width.* Observing that $S_{\mathbf{p}} \approx 0$ when $\mathbf{p} \neq 0$ we may write

$$E^{(1)} = E_a + C + \sum_{\mathbf{p}} e^{i(\mathbf{kp})} A(\mathbf{p}). \qquad (18.30)$$

Consider a simple cubic lattice as the first example. Every atom has six nearest neighbours with the co-ordinates

$$\mathbf{p} = a\left\{\begin{matrix} (1, & 0, & 0); & (-1, & 0, & 0) \\ (0, & 1, & 0); & (0, & -1, & 0) \\ (0, & 0, & 1); & (0, & 0, & -1) \end{matrix}\right\}. \qquad (18.31)$$

Assuming exchange energy to be isotropic $[A(\mathbf{p}) = A(|\mathbf{p}|) = A]$ and neglecting electron exchange between more distant atoms we obtain

$$E^{(1)} = E_a + C + 2A(\cos k_x a + \cos k_y a + \cos k_z a) = E(\mathbf{k}). \quad (18.32)$$

Energy depends quasicontinuously on the wave vector **k** and changes from E_{min} to E_{max}

$$\begin{aligned} E_{min} &= E_a + C - 6|A|, \\ E_{max} &= E_a + C + 6|A|. \end{aligned} \tag{18.33}$$

Thus, *as a result of atomic interaction the energy level of an isolated atom E_a drops by the amount C and splits into a band $12|A|$ wide (for a simple cubic lattice). Energy bands are separated by forbidden energy intervals—forbidden bands*. The energy band width depends on the exchange energy value A. But the exchange energy itself is dependent on the area of overlapping of the wave functions: the more the atomic wave functions overlap, the greater will the exchange energy be. It follows from here that energy levels corresponding to inner electron shells do not split so intensely as those of the outer shells since the inner shell electrons are localized in smaller areas of space. Figure 23 schematically shows the formation of bands from atomic levels.

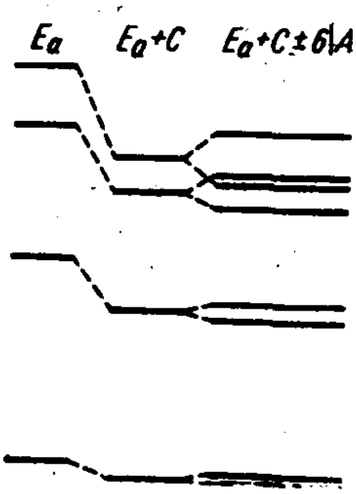


Fig. 23. The formation of energy bands out of energy levels when the atoms are brought together

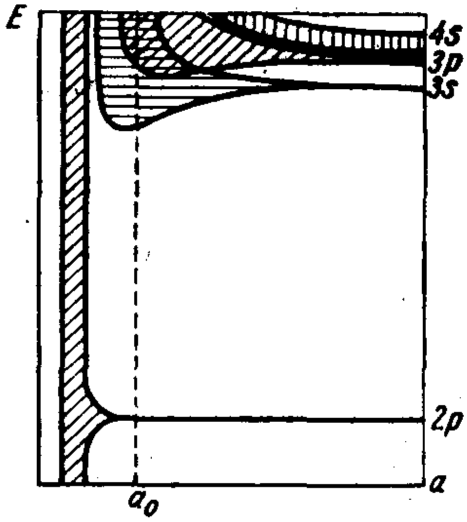


Fig. 24. The formation of the energy bands when the atoms are brought together

It is seen from Fig. 23 that *as energy increases, the bands become wider and the gaps narrower. The higher the level, the lower it drops and the wider it spreads.*

Figure 24 schematically shows how energy bands are formed from some energy levels as atoms are brought together, i.e. as the lattice period is changed. a_0 is the lattice period of the real crystal.

The interatomic distances for the tetrahedral and the rhombic systems are different for different crystallographic directions. Because of this the exchange integral, too, may depend on direction.

Representing the co-ordinates of neighbouring atoms p in the form

$$\mathbf{p}=\left\{\begin{matrix} (a_1, & 0, & 0); & (-a_1, & 0, & 0) \\ (0, & a_2, & 0); & (0, & -a_2, & 0) \\ (0, & 0, & a_3); & (0, & 0, & -a_3) \end{matrix}\right\}. \qquad (18.34)$$

and denoting exchange energy values corresponding to these directions by A_1, A_2, A_3 we may write

$$E(\mathbf{k})=E_a+C+2A_1\cos k_xa_1+2A_2\cos k_ya_2+2A_3\cos k_za_3. \qquad (18.35)$$

Find the extremal points assuming $\frac{dE(\mathbf{k})}{d\mathbf{k}}=0$. It is easily seen that energy extrema are located at points $k_{0i}=0$ and $k_{0i}=\pm\frac{\pi}{a_i}$, i.e. in the centre or in the apexes of the Brillouin zone.

Expanding $E(\mathbf{k})$ into the Taylor series around the points $\mathbf{k}_0$ we obtain the generalized expression

$$E(\mathbf{k})=E_{extr}+\frac{1}{2}\frac{d^2E}{d\mathbf{k}^2}\cdot(\mathbf{k}-\mathbf{k}_0)^2+\ldots, \qquad (18.36)$$

or

$$E(\mathbf{k})=E_0+C+2\sum_{i=1}^{3}A_i+\frac{1}{2}\sum_{i=1}^{3}\frac{\hbar^2}{m_i}k_i^2\,(k_{0i}=0) \qquad (18.37)$$

and

$$E(\mathbf{k})=E_0+C-2\sum_{i=1}^{3}A_i+\frac{1}{2}\sum_{i=1}^{3}\frac{\hbar^2}{m_i}\left(k_i\pm\frac{\pi}{a_i}\right)^2\left(k_{0i}=\pm\frac{\pi}{a}\right). \qquad (18.38)$$

For the effective mass $\mathbf{m}^*$ we have the expression

$$\mathbf{m}^{*-1}=\pm\begin{pmatrix} \frac{2A_1a_1^2}{\hbar^2} & 0 & 0 \\ 0 & \frac{2A_2a_2^2}{\hbar^2} & 0 \\ 0 & 0 & \frac{2A_3a_3^2}{\hbar^2} \end{pmatrix} \qquad (18.39)$$

where minus corresponds to the point $\mathbf{k}_0=0$, and plus—to the points $\left(\pm\frac{\pi}{a_1};\pm\frac{\pi}{a_2};\ \pm\frac{\pi}{a_3}\right)=\pm\pi(b_1, b_2, b_3)$.

The sign of the effective mass is determined by the exchange energy sign. When $A_i>0$ there is an energy maximum in the centre of the Brillouin zone. When $A_i<0$ the energy maxima are in the apexes and the minimum—in the centre of the Brillouin zone.

Energy is actually a quadratic function of $(\mathbf{k}-\mathbf{k}_0)$ periodic with a period $2\pi\mathbf{b}$. This follows directly from the analytical expression for $E(\mathbf{k})$.

The effective mass tensor components in the maximum and the minimum differ only in sign, their moduli being equal. In other words, for a definite zone the shape of constant-energy surfaces in the vicinity of a maximum or a minimum is identical. *The effective mass tensor components are inversely proportional to exchange energy*:

$$m_i = \pm \frac{\hbar^2}{2a_i^2 A_i}. \tag{18.40}$$

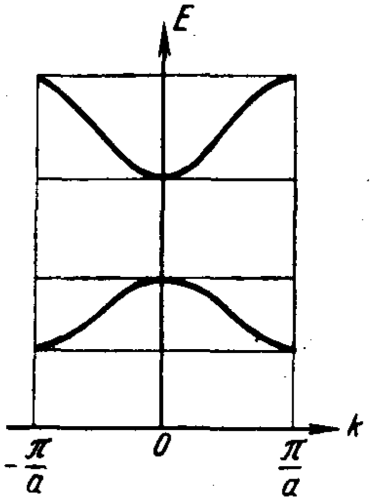


Fig. 25. The relation between the energy and the wave vector for $A > 0$ and $A < 0$

Since energy band width is determined by exchange energy, we may say that the *effective mass is inversely proportional to band width*. This means that *the higher the energy band, the smaller the effective mass of its charge carriers*.

Constant-energy surfaces in the vicinity of the extrema are ellipsoids. For instance, for $A_i < 0$ $(\mathbf{k}_0 = 0)$

$$E(\mathbf{k}) = E_{min} + \frac{\hbar^2}{2}\left(\frac{\mathbf{k}_x^2}{m_1} + \frac{\mathbf{k}_y^2}{m_2} + \frac{\mathbf{k}_z^2}{m_3}\right). \tag{18.41}$$

In this case the number of maxima will be eight, corresponding to the number of cube's apexes. However, since only 1/8 of each constant-energy surface built around the apex of the cube belongs to the first Brillouin zone, there is only one full ellipsoid to the first zone. This fact is of great importance for calculating energy-band states density.

For a simple cubic lattice exchange energy is isotropic: $A(\mathbf{p}) = A(|\mathbf{p}|)$, and therefore effective mass is a scalar

$$m^* = \pm \frac{\hbar^2}{2a^2 A}. \tag{18.42}$$

Constant-energy surfaces near an extremum are spheres. Figure 25 shows the depcndence $E(\mathbf{k})$. The form of constant-energy surfaces is shown in Fig. 17.

Let us now discuss the problem of the forbidden band width. We define the forbidden band width ΔE_0 between the nth and the $(n+1)$th energy bands as the minimum distance between them on the energy scale:

$$\Delta E_0 = E_{(n+1)\,min} - E_{n\,max}. \tag{18.43}$$

Substituting expression (18.37) and (18.38) for energy extrema in the two bands we obtain

$$\Delta E_0 = \left[E_{a\,(n+1)} + C_{(n+1)} - 2\sum_{i=1}^{3} |A_{i\,(n+1)}| \right] - \left[E_{an} + C_n + 2\sum_{i=1}^{3} |A_{in}| \right] = [E_{a\,(n+1)} - E_n] + [C_{(n+1)} - C_n] - 2\sum_{i=1}^{3} (|A_{i\,(n+1)}| + |A_{in}|). \quad (18.44)$$

This result is quite instructive. For the sake of simplicity we will assume that $C_{n+1} \approx C_n$. Then the forbidden band width will be narrower than the distance between the levels $E_{a\,(n+1)} - E_{an}$ by the sum of the neighbouring energy bands' half-widths. Since with the increase in energy the distances between energy levels decrease, and the energy band width, on the other hand, increases, this *results in narrowing the forbidden band width*. This result is basic for explaining the dependence of the forbidden band width on the atomic number of the element belonging to the same group. For instance, the width of the forbidden band between the valence band and the nearest free band for diamond, silicon, germanium and (grey) tin should decrease in the same order. This is fully substantiated by experiment.

Of great importance for the understanding of numerous physical phenomena is the sign of A_i for two neighbouring bands. If the signs of $A_{i\,(n+1)}$ and A_{in} are the same this means that energy maximum and minimum for these bands are located at different points of the Brillouin zone (in the centre and in the apexes of the zone). If, on the other hand, the signs of $A_{i\,(n+1)}$ and A_{in} are different the minimum and maximum are located at the same points of the Brillouin zone (Fig. 25). Each energy band has a corresponding Brillouin zone which we assume to be superimposed one on the other.

The cosine dependence of energy $E(\mathbf{k})$ for a simple cubic lattice was obtained on the assumption that exchange energy $A(1, 0, 0)$ exists only between nearest atoms. If more distant interaction is taken into account the expression for $E(\mathbf{k})$ will become more complicated; however, main qualitative deductions will remain unaltered.

Calculate the exchange energy $A(1, 1, 0)$ of the atom at the origin $(0, 0, 0)$ with atoms having the co-ordinates $a(1, 1, 0)$, $(1, 0, 1)$, etc. The expression for the energy will in this case take the form:

$$E(\mathbf{k}) = E_a + C + 2A(1,\ 0,\ 0)(\cos k_x a + \cos k_y a + \cos k_z a) +$$
$$+ 2A(1,\ 1,\ 0)[\cos(k_x + k_y)a + \cos(k_x - k_y)a + \cos(k_x + k_z)a +$$
$$+ \cos(k_x - k_z)a + \cos(k_y + k_z)a + \cos(k_y - k_z)a]. \quad (18.45)$$

If $|A(110)| \ll |A(100)|$ the change in the nature of constant-energy surfaces will be irrelevant. If, on the other hand, $|A(100)| \sim |A(110)|$ the changes both in the location of the extrema and in the shape of constant-energy surfaces will be marked.

Let us discuss the dependence of energy on quasimomentum for a volume-centered lattice taking account only of nearest neighbouring interaction between the atoms. If the edge of the cube is denoted by a, and the origin of co-ordinates is made to coincide with the atom in the centre of the cube, the co-ordinates of atoms in the apexes of the cube will be

$$\mathbf{p} = \frac{a}{2}\left\{\begin{array}{cc} (1,\ 1,\ 1); & (-1,\ -1,\ -1) \\ (-1,\ 1,\ 1); & (1,\ -1,\ -1) \\ (1,\ -1,\ 1); & (-1,\ 1,\ -1) \\ (1,\ 1,\ -1); & (-1,\ -1,\ 1) \end{array}\right\}. \quad (18.46)$$

Substituting (18.46) into (18.30) we obtain

$$E(\kappa) = E_0 + C + 2A\left[\cos(k_x + k_y + k_z)\frac{a}{2} + \cos(-k_x + k_y + k_z)\frac{a}{2} +\right.$$
$$\left. + \cos(k_x - k_y + k_z)\frac{a}{2} + \cos(k_x + k_y - k_z)\frac{a}{2}\right] =$$
$$= E_a + C + 8A\cos\frac{k_x a}{2}\cdot\cos\frac{k_y a}{2}\cdot\cos\frac{k_z a}{2}. \quad (18.47)$$

As can be seen from the expression (18.47) for $E(\mathbf{k})$, extremal energy values are subject to the condition that cosine moduli should be equal to unity and

$$\begin{aligned} E_{min} &= E_a + C - 8|A|, \\ E_{max} &= E_a + C + 8|A|. \end{aligned} \quad (18.48)$$

This gives the value 16 $|A|$ for the energy band width.

If a cube with the edge $\frac{4\pi}{a}$ is constructed in space it will contain 27 extremal points—one in the centre of the cube, six in the centres of the faces, eight in apexes and twelve in the middle of the edges. However, the volume thus defined is larger than the volume of the first Brillouin zone which has the shape of a dodecahedron.

Find the dependence of energy on the wave vector for a face-centered lattice. Placing the origin of co-ordinates in the apex of the cube with the edge a we write out the co-ordinates of the twelve nearest atoms:

$$\mathbf{p}=\frac{a}{2}\left\{\begin{array}{llllllll} (1, & 1, & 0); & (1, & 0, & 1); & (0, & 1, & 1) \\ (-1, & 1, & 0); & (1, & 0, & -1); & (0, & -1, & 1) \\ (1, & -1, & 0); & (-1, & 0, & 1); & (0, & 1, & -1) \\ (-1, & -1, & 0); & (-1, & 0, & -1); & (0, & -1, & -1) \end{array}\right\}. \quad (18.49)$$

Presuming as before the exchange energy to be isotropic we may write

$$\begin{aligned} E(\mathbf{k}) = E_a + C + 2A\Big[&\cos(k_x+k_y)\frac{a}{2}+\cos(k_x-k_y)\frac{a}{2}+ \\ &+\cos(k_x+k_z)\frac{a}{2}+\cos(k_x-k_z)\frac{a}{2}+\cos(k_y+k_z)\frac{a}{2}+ \\ &+\cos(k_y-k_z)\frac{a}{2}\Big] = E_a+C+4A\Big[\cos\frac{ak_x}{2}\cdot\cos\frac{ak_y}{2}+ \\ &+\cos\frac{ak_x}{2}\cdot\cos\frac{ak_z}{2}+\cos\frac{ak_y}{2}\cos\frac{ak_z}{2}\Big]. \end{aligned} \quad (18.50)$$

Suppose, to be definite, that $A>0$. In this case there is an energy maximum in the centre of the Brillouin zone equal to

$$E_{max}=E_a+C+12A. \quad (18.51)$$

The energy minimum cannot be equal to $E_a+C-12A$ for in this case binary cosine products should be equal to -1, and this is impossible. Writing out the conditions for an extremum for (18.51) we find that

$$E_{min}=E_a+C-4A, \quad (18.52)$$

i.e. energy band width is $16\,|A|$ and not $24\,|A|$.

In conclusion of the section let us stop to consider a point connected with the degeneracy of atomic levels. The s-level of an atom is singular, all others are degenerate. The degeneracy factor (with no account of spin) is equal to $g=(2l+1)$, where l is the orbital quantum number. Accordingly, p-levels are triple degenerate, d-levels—quintuple, etc. Interaction of the orbital and the spin magnetic moments of the electron results in so-called fine structure.

But if atomic energy levels are degenerate, in constructing electron wave functions for a crystal account should be taken of the degeneracy of atomic states:

$$\psi^0(\mathbf{r})=\sum_{\mathbf{n},\,\alpha} e^{i(\mathbf{kn})}c_\alpha\psi_{a\alpha}(\mathbf{r}-\mathbf{n}), \quad (18.53)$$

where c_α is an unknown coefficient and α assumes g different values. Substituting (18.53) into (18.19) we may calculate the energy E. The distinction of the last case from the one discussed above is that now integrals calculated with the aid of different wave functions are different. In consequence, degeneracy may be partially or totally removed and *three different energy branches E* (k) *obtained for the p-band.*

Summary of Sec. 18

1. The theory of the quasibound electron accepts the Hamiltonian for the electron in an isolated atom as zero approximation for the lattice field Hamiltonian. The perturbation operator is made up of the energy of interatomic interaction and of the energy of the electron in the field of all other atoms (except the nearest to it). The electron wave function in a crystal is a linear combination of atomic wave functions which satisfies the translational condition.

2. The "perturbation", i.e. the interaction of the electron belonging to a given atom with all other atoms (with their nuclei and electron shells) results in a significant change of atomic electron energy levels E_a: they are lowered and split into an energy band.

3. The drop in the level E_a is the greater, the higher is this level in an isolated atom, since the dimensions of the electron cloud increase with the increase in energy E_a.

4. If the atoms of a crystal are brought infinitely close together forces of repulsion will arise between atomic electron shells, the potential energy W (r) will turn positive and will continue to increase rapidly causing a rise in the energy E_a ($C > 0$).

5. Energy band width is proportional to the exchange integral which increases with the increase in E_a. Therefore, the higher is the energy band, the wider it is (Fig. 23).

6. The forbidden band is the narrower the higher are the corresponding bands. The decrease in the forbidden band width with the increase in energy is the result of the decrease in separation between respective energy levels and of the increase in energy band width. For high energy values energy bands may superimpose, or overlap (Fig. 23).

7. Effective electron mass is inversely proportional to the exchange integral and, by force of this, to the band width. Since the increase in band width, all other conditions being equal, results in a decrease of the forbidden band width, effective electron mass may be expected to be smaller in materials with narrower forbidden bands.

8. Actions which are accompanied by the change in interatomic distances (heating, compression or extension) result in the changes

in wave function overlapping and in corresponding changes in the exchange integral, band width, effective mass and forbidden band width. This, in turn, leads to changes in the physical properties of semiconductors which depend on the forbidden band width or on the effective mass.

9. The band structure of the energy spectrum follows both from the quasifree and the quasibound electron theories. The results of the former are more accurate for high energies while the latter is more valid for low energies.

19. EFFECTIVE MASS METHOD. INFLUENCE OF EXTERNAL FIELDS ON ENERGY SPECTRUM OF A CRYSTAL

Consider now the solution of the Schrödinger equation for the case when an external field $V(\mathbf{r})$ is applied to the crystal:

$$\hat{\mathbf{H}}\psi = \left[-\frac{\hbar^2}{2m}\Delta + U(\mathbf{r}) + V(\mathbf{r})\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}). \qquad (19.1)$$

To solve this equation the field $U(\mathbf{r})$ which is unknown to us should be known. There is, however, a convenient and a sufficiently accurate method of solving equation (19.1) which became known under the name of the effective mass method. To understand it let us again consider the solution of the Schrödinger equation for an ideal crystal. As we know, the electron wave function for an ideal crystal is the Bloch function $\psi_k(\mathbf{r})$, and electron energy in the vicinity of the extremum $\mathbf{k}_0$ is a quadratic function of quasimomentum or of the wave vector:

$$\hat{\mathbf{H}}_0\psi_{\mathbf{k}}(\mathbf{r}) = \left[-\frac{\hbar^2}{2m}\Delta + U(\mathbf{r})\right]\psi_{\mathbf{k}}(\mathbf{r}) = E(\mathbf{k})\psi_{\mathbf{k}}(\mathbf{r});$$

$$E(\mathbf{k}) = E(\mathbf{k}_0) + \frac{\hbar^2}{2m^*}(\mathbf{k}-\mathbf{k}_0)^2. \qquad (19.2)$$

Consider the Hamiltonian $\hat{\hat{\mathbf{H}}}_0$:

$$\hat{\hat{\mathbf{H}}}_0 = -\frac{\hbar^2}{2m^*}\Delta \qquad (19.3)$$

which is the Hamiltonian of a free particle with a mass equal to the effective electron mass. Solve the Schrödinger equation:

$$\hat{\hat{\mathbf{H}}}_0\tilde{\psi}_0 = \tilde{E}_0\tilde{\psi}_0. \qquad (19.4)$$

Substituting the expression for $\hat{\hat{\mathbf{H}}}_0$ we may write:

$$\sum_{i=1}^{3}\frac{1}{m_i}\frac{\partial^2}{\partial x_i^2}\tilde{\psi}_0 + \frac{2E_0}{\hbar^2}\tilde{\psi}_0 = 0. \qquad (19.5)$$

By direct substitution into this equation we can make sure that $\tilde{\psi}_0$ in the form of

$$\tilde{\psi}_0 = Ae^{i(\alpha x+\beta y+\gamma z)} \tag{19.6}$$

is a solution of equations (19.4) and (19.5). At the same time

$$\tilde{E}_0 = \frac{\hbar^2}{2}\left(\frac{\alpha^2}{m_1}+\frac{\beta^2}{m_2}+\frac{\gamma^2}{m_3}\right). \tag{19.7}$$

Comparing (19.7) and (19.2) we conclude that equation (19.4) will have the same energy spectrum as the equation (19.2) if α, β, γ are chosen in the form

$$\alpha = k_x - k_{0x};\quad \beta = k_y - k_{0y};\quad \gamma = k_z - k_{0z}, \tag{19.8}$$

and if the origin of the energy scale is made to coincide with the extremal point, i.e. if $E(\mathbf{k}_0)$ is made to equal 0. Generally, the Hamiltonian

$$\hat{\mathrm{H}}_0 = -\frac{\hbar^2}{2m^*}\Delta + E(\mathbf{k}_0) \tag{19.9}$$

has the same energy spectrum as the lattice field Hamiltonian in the vicinity of an extremum. Its eigenfunctions are of the form

$$\tilde{\psi}_0(\mathbf{r}) = \frac{1}{\sqrt{G}}e^{i(\mathbf{k}-\mathbf{k}_0,\,\mathbf{r})}. \tag{19.10}$$

Hence, *the motion of a particle in a crystal is analogous to the motion of a free particle. The difference between a really free particle and a particle in a crystal lies in their masses: in the crystal the effective mass* $\mathbf{m}^*$ *takes the place of the mass of a free particle.*

When external fields $V(\mathbf{r})$ are applied the Schröndinger equation (19.1) may be written in the form

$$\left[-\frac{\hbar^2}{2\mathrm{m}^*}\Delta + V(\mathbf{r})\right]\psi(\mathbf{r}) = E\psi(\mathbf{r}). \tag{19.11}$$

Having solved this equation we will find the wave function $\psi(\mathbf{r})$ and the energy E of the electron in a crystal in an external field. *The solution of problems based on equation* (19.11) *became known as the effective mass method. As compared with the general equation* (19.1) *it has the advantage that instead of the unknown lattice potential field* $U(\mathbf{r})$ *we may make use of experimentally determined effective mass* $\mathbf{m}^*$ *of the particle.* Its disadvantage is its approximate nature since the method makes sense only for states in the vicinity of the extrema.

With the aid of Wannier functions it can be demonstrated more rigorously that the equation (19.1) does not result in appreciable changes in the energy spectrum in case of smoothly changing external fields, and that the wave function is actually of the form (19.10).

To determine the changes in the energy spectrum of an ideal crystal caused by an external field we shall make use of the quantum equation of motion instead of the usual perturbation theory.

Consider the Hamiltonian of a crystal in an external field (19.1). *Quantum-mechanical equation of motion for an arbitrary physical quantity, not explicitly dependent on time and having the operator* $\hat{\mathbf{L}}$ *is of the form*

$$\frac{d\hat{\mathbf{L}}}{dt} = [\hat{\mathbf{H}}, \hat{\mathbf{L}}] = \frac{1}{i\hbar}(\hat{\mathbf{L}}\hat{\mathbf{H}} - \hat{\mathbf{H}}\hat{\mathbf{L}}). \qquad (19.12)$$

Apply this equation to the ideal crystal Hamiltonian

$$\hat{\mathbf{H}}_0 = -\frac{\hbar^2}{2m}\Delta + U(\mathbf{r}).$$

$$\frac{d\hat{\mathbf{H}}_0}{dt} = [\hat{\mathbf{H}}, \hat{\mathbf{H}}_0] = [\hat{\mathbf{V}}, \hat{\mathbf{H}}_0] = [\hat{\mathbf{V}}, \hat{\mathbf{T}}] =$$

$$= \frac{1}{i\hbar}\left(-\frac{\hbar^2}{2m}\right)\{\Delta\hat{\mathbf{V}} - \hat{\mathbf{V}}\Delta\} = \frac{i\hbar}{2m}\{(\Delta V) + 2(\nabla V, \nabla)\} \qquad (19.13)$$

Take the case of fields slowly changing in space. For such fields ΔV may be neglected as compared with the second addend. Taking into account that $(\nabla V) = -\mathbf{F}_a$, and that

$$-\frac{i\hbar}{m}\nabla = \frac{\hat{\mathbf{p}}}{m} = \hat{\mathbf{v}} = \frac{d\hat{\mathbf{r}}}{dt} \qquad (19.14)$$

we write

$$\frac{d\hat{\mathbf{H}}_0}{dt} = \left(\hat{\mathbf{F}}_a, \frac{d\hat{\mathbf{r}}}{dt}\right). \qquad (19.15)$$

Average this expression over the initial states with the wave function $\psi_{\mathbf{k}}(\mathbf{r})$. Supposing that $\mathbf{F}_a$ is independent of the co-ordinate we can take $\mathbf{F}_a$ out of the integral sign:

$$\left\langle \frac{d\hat{\mathbf{H}}_0}{dt} \right\rangle = \frac{d}{dt}\int \psi_{\mathbf{k}}^*(\mathbf{r})\,\hat{\mathbf{H}}_0\psi_{\mathbf{k}}(\mathbf{r})\,d\tau = \frac{dE_0(\mathbf{k})}{dt} =$$

$$= \int \psi_{\mathbf{k}}^*(\mathbf{r})\,(\hat{\mathbf{F}}_a \cdot \hat{\mathbf{v}})\,\psi_{\mathbf{k}}(\mathbf{r})\,d\tau = (\mathbf{F}_a \cdot \langle \mathbf{v} \rangle). \qquad (19.16)$$

Equation (19.16) is classical and it shows that *"total" energy* $E_0(\mathbf{k})$ *changes in time as a result of work performed by the external force* $\mathbf{F}_a$. This equation can be derived from the equation of motion

$$\frac{d\mathbf{P}}{dt} = \mathbf{F}_a = -\nabla V.. \qquad (19.17)$$

Multiply (19.17) by velocity $\mathbf{v}=\frac{d\mathbf{r}}{dt}$

$$\left(\mathbf{v}\cdot\frac{d\mathbf{P}}{dt}\right)=\left(\frac{\mathbf{P}-\mathbf{P}_0}{m^*}\cdot\frac{d\mathbf{P}}{dt},\right)=\frac{d}{dt}\cdot\frac{(\mathbf{P}-\mathbf{P}_0)^2}{2m^*}=$$
$$=\frac{d}{dt}E_0(\mathbf{P})=-\frac{dV}{d\mathbf{r}}\frac{d\mathbf{r}}{dt}=-\frac{dV}{dt} \qquad (19.18)$$

or

$$E_0(\mathbf{P})+V(\mathbf{r})=H_0(\mathbf{P},\mathbf{r})=\text{const.} \qquad (19.19)$$

$H_0(\mathbf{P},\mathbf{r})$ in our notation is the time-constant Hamilton function of the electron in a crystal to which an external field $V(\mathbf{r})$ has been applied. Equation (19.11) exactly corresponds to the equation (19.19), and the effective mass method is thereby rigorously substantiated by quantum mechanics.

Note that total energy H_0 of the electron moving in a crystal under the influence of an external field is conserved; at the same time $E_0(\mathbf{P})$ and $V(\mathbf{r})$ both change but in opposite directions. $V(\mathrm{r})$ may change within arbitrary limits, $E_0(\mathbf{P})$, on the other hand, only within the energy band: from E_{min} to E_{max}. But this leads to the conclusion that the change in $V(\mathrm{r})$ should be limited to the same interval. To prove it take the differentials of the left and right members of the equation (19.19) to obtain

$$dE_0+dV=0, \quad \text{or} \quad \delta E_0+\delta V=0. \qquad (19.20)$$

Setting $\delta E_0=E_{max}-E_{min}$ we obtain

$$\delta V=-\delta E_0=E_{min}-E_{max}. \qquad (19.21)$$

From (19.21) it follows also that *in a uniform electric field the electron should oscillate periodically along the field in an interval δr long*

$$\delta V=-(\delta\mathbf{r},\mathbf{F}_a)=-(E_{max}-E_{min}) \qquad (19.22)$$

whence

$$\delta r=\frac{E_{max}-E_{min}}{F_a}=\frac{E_{max}-E_{min}}{e\,|\,\mathbf{E}\,|}. \qquad (19.23)$$

Consider again equation (19.19). Fix the value $\mathbf{P}=\mathbf{P}'$ and consider all values the Hamilton function H_0 may assume when the co-ordinate is changed. Since $E_0(\mathbf{P})$ is independent of the co-ordinate, the dependence of H_0 on r will be determined by the external field $V(\mathbf{r})$ (Fig. 26).

Figure 26*b* may be regarded as representing the dependence of the energy level $E_0(\mathbf{P}')$ on the coordinate.

If now the energy $E_0(\mathbf{P}')$ is made to run through all possible values the graphical dependence of energy band position on the co-ordinate will be obtained (Fig. 27). *The graph $H_0(\mathbf{r})$ may be interpre-*

ted as showing energy band bending under the influence of external fields. This interpretation is very convenient since it enables both "free" electron motion and motion with scattering to be considered.

Consider the "free" motion described by equation (19.19). The total energy of the electron moving in an external field $V(\mathbf{r})$ is conserved: $H_0 = \text{const}$. This is shown in Fig. 27 by a straight line parallel to the x-axis. We see from here that moving periodically

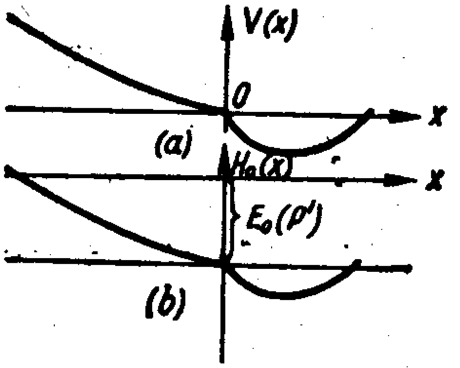


Fig. 26. The dependence of the potential energy and the Hamilton function on the coordinates

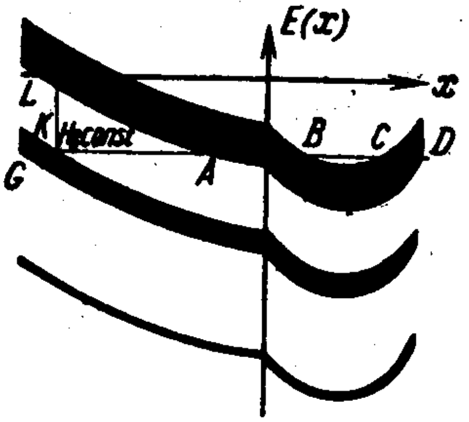


Fig. 27. The bending of the energy bands in external fields

in the crystal the electron passes from one energy level to other levels of the band. Electron with the same total energy $H_0 = \text{const}$ can oscillate between the points $A-B$, $C-D$ and $G-K$ belonging to two bands. In accordance with (19.23) the area δr will be the smaller the greater is the inclination of the bands (compare $C-D$ with $A-B$) and the narrower the energy band (compare $A-B$ with $G-K$). The electron can, however, pass from the points of interval $G-K$ to the points of interval $A-B$ and thereby from the second to the third (i.e. "higher") energy band. This transition is possible via the tunnel effect—the transition from point K to point A entails passing through a triangular potential barrier KLA. Quantum mechanics establishes that the probability of tunnelling depends exponentially on the width and height of the potential barrier. The height of the potential barrier KL coincides with the forbidden band width ΔE_0. The barrier width KA, in turn, depends on KL and on the band inclination which is a function of the force $\mathbf{F}_a$, i.e. of field intensity $\mathbf{E}$:

$$KA = \frac{KL}{F_a} = \frac{\Delta E_0}{|e\mathbf{E}|}\,. \tag{19.24}$$

However, since the barrier height enters the power of the exponent in the form of its square root, we arrive at the conclusion that the *permeability of a triangular barrier is proportional to*

$\exp\left\{\gamma\frac{\Delta E_0^{3/2}}{|\mathbf{E}|}\right\}$, *i.e. that band-to-band tunnelling probability increases exponentially with the increase in the external field intensity* **E** *which causes the bands to incline.* Since the transition probabilities from point K to point A and vice-versa are equal *the predominant transition shall take place from the band with the higher electron concentration. Rapid increase in the number of free electrons in the upper free band at the expense of the lower filled band caused by electric fields is known as the Zener effect.*

Tunnelling takes place in the interval BC, as well. Consider now "unfree" motion of the electron when it transmits energy received n the field $V(\mathbf{r})$, for example, to the lattice. In this case *the lectron remains on the same energy level and can moιe to any ditance along the crystal.*

Summary of Sec. 19

1. The equation used to describe electron motion in a crystal to which an external potential field $V(\mathbf{r})$ has been applied is

$$\left[-\frac{\hbar^2}{2m^*}\Delta+V(\mathbf{r})\right]\psi(\mathbf{r})=E\psi(\mathbf{r}). \tag{19.1s}$$

This is the equation to which equation (19.1) is reduced when the periodic field operator $\hat{\mathbf{H}}_0=-\frac{\hbar^2}{2m}\Delta+U(\mathbf{r})$ is replaced by the operator $\hat{\mathbf{H}}_0=-\frac{\hbar^2}{2m^*}\Delta$ of a free particle whose mass is equal to the effective mass of the electron in the solid. This method of solving the Schrödinger equatioń became known as the effective mass method.

2. The effective mass method is valid for non-perturbed states with energies close to extremal values and for slowly changing applied fields. It enables the unknown quantity $U(\mathbf{r})$ to be excluded from the equation (19.1) by the introduction of the effective mass $\mathbf{m}^*$ which can be measured experimentally.

3. External fields $V(\mathbf{r})$ bend energy bands of an ideal crystal. Energy levels are raised in the parts of the crystal where $V(\mathbf{r})>0$ and lowered where $V(\mathbf{r})<0$. The displacement of all bands is the same and equals $V(\mathbf{r})$ at every point.

4. Band inclination in strong electric fields results in the tunnelling of electrons from the "lower" to the "upper" bands. This phenomenon is known by the name of the Zener effect.

20. LOCALIZED STATES

Suppose now that a sufficiently strong perturbation W *localized in a small part of the crystal with the "centre" at point* $\mathbf{R}$ is applied to the crystal. Denote this by $W(\mathbf{r}-\mathbf{R})$. The Schrödinger equation assumes the form:

$$\left[-\frac{\hbar^2}{2m}\Delta+U(\mathbf{r})+W(\mathbf{r})\right]\psi(\mathbf{r})=E\psi(\mathbf{r}), \qquad (20.1)$$

or

$$[\hat{\mathrm{H}}_0+\hat{\mathrm{W}}]\psi(\mathbf{r})=E\psi(\mathbf{r}). \qquad (20.2)$$

Expanding the solution sought $\psi(\mathbf{r})$ into a series of Bloch functions $\psi_{\mathbf{k}}(\mathbf{r})$

$$\psi(\mathbf{r})=\sum_{\mathbf{k}} c_{\mathbf{k}}\psi_{\mathbf{k}}(\mathbf{r}) \qquad (20.3)$$

and substituting (20.3) into equation (20.2), we obtain

$$\sum_{\mathbf{k}} c_{\mathbf{k}}[E_0(\mathbf{k})+\hat{\mathrm{W}}]\psi_{\mathbf{k}}(\mathbf{r})=E\sum_{\mathbf{k}} c_{\mathbf{k}}\psi_{\mathbf{k}}(\mathbf{r}). \qquad (20.4)$$

Next we multiply the equation by $\psi_{\mathbf{k}'}(\mathbf{r})$, integrate it over the entire crystal and simplify it by writing it out in the usual matrix form:

$$c_{\mathbf{k}'}E_0(\mathbf{k}')+\sum_{\mathbf{k}} W_{\mathbf{k}'\mathbf{k}}c_{\mathbf{k}}=Ec_{\mathbf{k}'}, \qquad (20.5)$$

or

$$[E-E_0(\mathbf{k}')]c_{\mathbf{k}'}-\sum_{\mathbf{k}} W_{\mathbf{k}'\mathbf{k}}c_{\mathbf{k}}=0. \qquad (20.6)$$

Instead of writing out the secular equation from which the allowed energy value E is to be obtained we will re-write the equation (20.6) taking account of the following condition imposed on $W(\mathbf{r}-\mathbf{R})$: the perturbation is small (or zero) everywhere except a small area V_R centred around $\mathbf{R}$. Therefore, when calculating the matrix element

$$W_{\mathbf{k}'\mathbf{k}}=\int\limits_{(V_R)}\psi^*_{\mathbf{k}'}(\mathbf{r})\,W(\mathbf{r}-\mathbf{R})\,\psi_{\mathbf{k}}(\mathbf{r})\,d\tau, \qquad (20.7)$$

one should integrate over a small volume containing the point $\mathbf{R}$. Making use of the mean value theorem set the values of the wave functions equal to those at some point $\mathbf{R}'$ of the volume V_R and take them out of the integral:

$$W_{\mathbf{k}'\mathbf{k}}=\psi^*_{\mathbf{k}'}(\mathbf{R}')\psi_{\mathbf{k}}(\mathbf{R}')\cdot\int W(\mathbf{r}-\mathbf{R})\,d\tau=\psi^*_{\mathbf{k}'}(\mathbf{R}')\psi_{\mathbf{k}}(\mathbf{R}')\,W_0V_R, \qquad (20.8)$$

where

$$W_0 = \frac{1}{V_R} \cdot \int\limits_{(V_R)} W(\mathbf{r}-\mathbf{R})\, d\tau \qquad (20.9)$$

is the *energy of perturbation averaged over the volume* V_R.

Substituting the expression for $W_{\mathbf{k}'\mathbf{k}}$ into the equation (20.6) we obtain

$$[E-E_0(\mathbf{k}')]\, c_{\mathbf{k}'} - \psi^*_{\mathbf{k}'}(\mathbf{R}')\, W_0 V_R \sum_{\mathbf{k}} c_{\mathbf{k}} \psi_{\mathbf{k}}(\mathbf{R}') = 0. \qquad (20.10)$$

But

$$\sum_{\mathbf{k}} c_{\mathbf{k}} \psi_{\mathbf{k}}(\mathbf{R}') = \psi(\mathbf{R}') \qquad (20.11)$$

does not depend on **k** and **k**′, and for this reason we may using (20.10) and (20.11) write the expression for $c_{\mathbf{k}'}$:

$$c_{\mathbf{k}'} = \frac{\psi_{\mathbf{k}'}(\mathbf{R}')\, W_0 V_R \psi(\mathbf{R}')}{E-E_0(\mathbf{k}')}. \qquad (20.12)$$

Now substitute the expressions for $c_{\mathbf{k}'}$ and $c_{\mathbf{k}}$ into the equation (20.10), cancel out $\psi_{\mathbf{k}'}$, $\psi(\mathbf{R}')$ and $W_0 V_R$ and write the equation in the form:

$$1 - \sum_{\mathbf{k}} \frac{\psi^*_{\mathbf{k}}(\mathbf{R}')\, \psi_{\mathbf{k}}(\mathbf{R}')\, W_0 V_R}{E-E_0(\mathbf{k})} = 0, \qquad (20.13)$$

or

$$\sum_{\mathbf{k}} \frac{\psi^*_{\mathbf{k}}(\mathbf{R}')\, \psi_{\mathbf{k}}(\mathbf{R}')}{E-E_0(\mathbf{k})} = \frac{1}{W_0 V_R}. \qquad (20.14)$$

We see from here that *energy eigenvalues of a perturbed system are related to the localized perturbation by means of* $W_0 V_R$ *and the values of wave functions at some point* **R**′ *near the point of perturbation.*

The equation (20.14) is an algebraic equation of the Nth degree in respect to E (where N is the number of the values of $\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_N$). Therefore it has N roots. Suppose $W_0 V_R \to 0$; in this case the right-hand side tends to infinity. Since $|\psi_{\mathbf{k}}(\mathbf{R}')|^2$ is a constant independent of $W_0 V_R$, for the equation (20.14) to be satisfied at least one of the addends in (20.14) should tend to infinity (or the denominator of one of the addends—to zero), say, for $\mathbf{k} = \mathbf{k}''$

$$E - E_0(\mathbf{k}'') \to 0, \qquad (20.15)$$

where **k**″ assumes one of the N possible values of **k**, and E—one of the allowed values of the energy of the unperturbed system $E_0(\mathbf{k}'')$. In other words, when the perturbation W tends to zero the solution of the "perturbed" problem, naturally, reduces to the solution of the "unperturbed" problem.

Consider another limiting case: $W_0V_R \rightarrow \pm\infty$. In this case the right-hand side of (20.14) tends to zero. Introduce for the sake of convenience the following notation:

$$\mathbf{k} = \mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_N; \quad E_0(\mathbf{k}_n) = a_n; \quad E = x;$$
$$|\psi_{\mathbf{k}_n}(\mathbf{R}')|^2 = A_n; \quad \frac{1}{W_0V_R} = B \tag{20.16}$$

and consider a function of two variables x and B

$$f(x, B) = \left[\sum_{n=1}^{N} \frac{A_n}{x-a_n} - B\right] \cdot \prod_{i=1}^{N} (x-a_i). \tag{20.17}$$

Allowed energy values E are determined by the condition $f(x, B) = 0$. But for $f(x, B) = 0$ the following conditions should be satisfied:

$$\left[\sum_{n=1}^{N'} \frac{A_n}{x-a_n} - B\right] = 0; \quad \prod_{i=1}^{N} (x-a_i) = 0. \tag{20.18}$$

The first determines the roots of the perturbed problem, the second — those of the unperturbed and will therefore be of no interest to us. Below we will suppose $\prod_{i=1}^{N} (x-a_i) \neq 0$. Represent $f(x, B)$ in another form:

$$f(x, B) = -B\prod_{i=1}^{N}(x-a_i) + \sum_{n}^{N} \prod_{i \neq n}^{N} (x-a_i) \cdot A_n =$$
$$= -Bx^N + \left[\sum_n A_n + B\sum_n a_n\right] x^{N-1} + \ldots +$$
$$+ \left[\sum_n \frac{A_n}{a_n} \prod_{i=1}^{N} a_i + B\prod_{i=1}^{N} a_i\right] x(-1)^{N+1}. \tag{20.19}$$

The function $f(x, B)$ is linear in B and is a polynomial of the Nth degree with respect to x; therefore for any $B \neq 0$ it has N roots.

When $x \rightarrow \infty$ the main part will be played by the term of the highest power, i.e. $-Bx^N$. For $f(x, B)$ to remain finite B must tend to zero or, vice versa, the conclusion may be drawn that if $B \rightarrow 0$, $x \rightarrow \infty$ approximately as B^{-1}. Hence $E \rightarrow \pm\infty$ when $W_0V_R \rightarrow \rightarrow \pm\infty$.

The question arises: do all the roots of (20.19) tend to infinity? Let $B=0$. Then

$$f(x,0)=\left(\sum_{n=1}^{N}A_n\right)x^{N-1}+\ldots+(-1)^{N+1}\prod_{l=1}^{N}a_l\sum_{n}\frac{A_n}{a_n}, \qquad (20.20)$$

i,e. for $B=0$ $f(x,0)$ is a polynomial of $(N-1)$th degree with respect to x and has $(N-1)$ roots which coincide with the roots of the unperturbed equation.

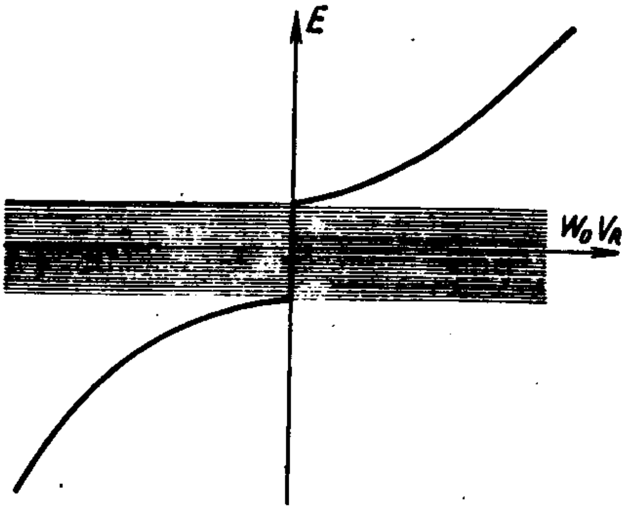


Fig. 28. The splitting of an energy level away from an energy band upon application of a localized perturbation to the crystal

We may therefore draw the following conclusion: *when a sufficiently intense localized perturbation is applied to a crystal one of the energy sublevels of the unperturbed system splits away from the band and either drops or rises. The position of the remaining* $(N-1)$ *levels remains practically unaltered.* When $W_0V_R>0$ the maximum energy level rises, and when $W_0V_R<0$ the minimum energy level drops (Fig. 28). This leads to the *creation of an allowed state inside the forbidden band.*

Consider the wave function of a state in the forbidden band. Let the energy of this state be $E_l>E(\mathbf{k})$. The wave function may be written in the form:

$$\psi_l(\mathbf{r})=\sum_{\mathbf{k}}c_{\mathbf{k}}\psi_{\mathbf{k}}(\mathbf{r})=\sum_{\mathbf{k}}\frac{\psi_{\mathbf{k}}^*(\mathbf{R}')W_0V_R\psi_l^*(\mathbf{R}')}{E_l-E_0(\mathbf{k})}\psi_{\mathbf{k}}(\mathbf{r})=$$
$$=\psi_l(\mathbf{R}')W_0V_R\sum_{\mathbf{k}}\frac{\psi_{\mathbf{k}}^*(\mathbf{R}')}{E_l-E_0(\mathbf{k})}\cdot\psi_{\mathbf{k}}(\mathbf{r}). \qquad (20.21)$$

All the functions $\psi_{\mathbf{k}}(\mathbf{r})$ take part in the formation of $\psi_l(\mathbf{r})$, the weight of each function being $\frac{\psi_{\mathbf{k}}^*(\mathbf{R}')}{E_l-E_0(\mathbf{k})}$. Find the limit of $\psi_l(\mathbf{r})$ when $W_0V_R\to\infty$.

In this case

$$\frac{E_l}{W_0 V_R} \approx 1, \quad E_l \gg E_0(\mathbf{k}) \tag{20.22}$$

and

$$\psi_l(\mathbf{r}) \approx \psi_l(\mathbf{R}') \sum_{\mathbf{k}} \psi_{\mathbf{k}}^*(\mathbf{R}') \psi_{\mathbf{k}}(\mathbf{r}) =$$

$$= \psi_l(\mathbf{R}') \sum_{\mathbf{k}} e^{-i(\mathbf{kR}')} \varphi_{\mathbf{k}}^*(\mathbf{R}') \psi_{\mathbf{k}}(\mathbf{r}). \tag{20.23}$$

If in the latter expression $\varphi_{\mathbf{k}}^*(\mathbf{R}')$ is replaced by some maximum quantity $\varphi^*(\mathbf{R}')$ independent of $\mathbf{k}$ it will be possible to assess the sum (20.23):

$$\psi_l(\mathbf{r}) \leqslant \psi_l(\mathbf{R}') \varphi^*(\mathbf{R}') \sum_{\mathbf{k}} e^{-i(\mathbf{kR}')} \psi_{\mathbf{k}}(\mathbf{r}). \tag{20.24}$$

Introduce the notation

$$\sum_{\mathbf{k}} e^{-i(\mathbf{kR}')} \psi_{\mathbf{k}}(\mathbf{r}) = \sqrt{N}\, \Phi(\mathbf{r}-\mathbf{R}'). \tag{20.25}$$

The function $\Phi(\mathbf{r}-\mathbf{R}')$ is a Wannier function. It is not equal to zero only in a small vicinity of the point $\mathbf{R}'$. In fact, $\Phi(\mathbf{r}-\mathbf{R}')$ may be represented in the form

$$\Phi(\mathbf{r}-\mathbf{R}') = \frac{1}{\sqrt{N}} \sum_{\mathbf{k}} e^{i(\mathbf{k},\, \mathbf{r}-\mathbf{R}')} \varphi_{\mathbf{k}}(\mathbf{r}) \cong$$

$$\cong \frac{1}{\sqrt{N}} \sum_{\mathbf{k}} e^{i(\mathbf{k},\, \mathbf{r}-\mathbf{R}')} \varphi_{\mathbf{k}}(\mathbf{r}-\mathbf{R}') =$$

$$= \frac{1}{\sqrt{N}} \sum_{\mathbf{k}} e^{i(\mathbf{k},\, \mathbf{r}')} \varphi_{\mathbf{k}}(\mathbf{r}') = \frac{1}{\sqrt{N}} \sum_{\mathbf{k}} \psi_{\mathbf{k}}(\mathbf{r}'). \tag{20.26}$$

But the sum of a great number of periodic functions (plane waves) is close to zero everywhere except the point $\mathbf{r}'=0$ and its close vicinity. Therefore the Wannier function is not equal to zero only in a small vicinity of the point $\mathbf{R}'$.

We see that the wave function $\psi_l(\mathbf{r})$ is proportional to the Wannier function $\Phi(\mathbf{r}-\mathbf{R}')$, i.e. $\psi_l(\mathbf{r})$ is not equal to zero only near $\mathbf{R}'$. Actually the wave packet of $\psi_l(\mathbf{r})$ will be distributed over a somewhat greater area than the Wannier function because of the dependence of $\varphi_{\mathbf{k}}^*(\mathbf{R}')$ on $\mathbf{k}$; however, the nature of the wave function will remain the same. This result is very important—the electron having the energy E_l lying inside the forbidden band is localized in a volume of the order of V_R. In other words, the *electron is localized within the area of perturbation.*

For a finite $W_0 V_R$ value the area of localization becomes larger: *the smaller the perturbation energy the greater the localization area.* This follows directly from the expression for $\psi_l(\mathbf{r})$: the closer is

E_l to $E_0(\mathbf{k})$ the greater is the weight of corresponding wave functions $\psi_k(\mathbf{r})$. When $W_0V_R \rightarrow 0$ the localized state function $\psi_l(\mathbf{r})$ is transformed into one of the functions $\psi_k(\mathbf{r})$ which describe a totally delocalized electron state.

This result is of fundamental importance to semiconductor physics.

Summary of Sec. 20

1. Local lattice field imperfections such as impurity atoms, vacancies and other point defects, and dislocations bring about the formation of allowed states connected with the perturbation area inside the forbidden band.

2. The electron wave function of such states is non-zero in approximately the same area where the perturbation exists. In other words, the electron is localized in the perturbation area.

3. The formation of the energy level corresponding to the localized state is the result of the splitting off of the extremal energy level and its transition to the forbidden band. If the perturbation is negative the energy level in the forbidden band will be formed as the result of the splitting off of the minimum energy level. If, on the other hand, the perturbation is positive it is the maximum energy level (of a different band) that goes over into the forbidden band. The perturbation existing around a vacancy may serve as an example of a positive perturbation. The stronger the perturbation, the greater the separation of the localized state from the corresponding band (see Fig. 28).

4 As the concentration of localized perturbation N_l increases, the mean distance between them decreases since it is equal to $N_l^{-1/3}$, and the wave functions of localized states may begin to overlap. This will bring about exchange interaction between localized states with the resultant splitting of the discrete energy level E_l into an energy band.

If localized perturbations are due to impurity atoms corresponding levels and band are termed impurity levels and impurity band. A sufficiently high concentration of impurity atoms is required to turn the impurity level into the impurity band.

21. ELEMENTARY THEORY OF IMPURITY STATES

Doping is the most important cause of impurity states formation in semiconductors. Let us now assess the position of an impurity level inside the forbidden band. The transition of an electron from an impurity level into an energy band is tantamount to the ionization of an impurity atom, therefore it is convenient to take the bottom of the band as the origin of the energy scale. The separation between the bottom of the band and the impurity level should be equal to the energy of ionization of the impurity atom. The ionization energy comprises from 4 to 5 eV for alcaline metals, to 24 eV for He; roughly, it may be taken equal to ten electron-volts. When an impurity atom is introduced into a semiconductor it interacts with the atoms of the matrix with the result that the binding energy of the electron with the impurity atom is decreased as compared to the binding energy of the electron in a free atom. Take the example of an atom of the group V element As in germanium. As was already mentioned, four valence electrons of the arsenic atom will take part in the formation of four two-electron bonds with the nearest germanium atoms. The fifth electron will not take part in the formation of co-valent bonds and will interact comparatively weakly with many germanium atoms. Its bond with the As atom will be weakened, and it will start moving in an orbit of great radius ,embracing tens or hundreds of the matrix atoms (if one speaks in terms of visual concepts of the Bohr theory). For this reason the interaction of the excess electron with the matrix atoms may formally be defined in terms of the dielectric permeability ε of the matrix. Accordingly, let the potential energy of the impurity atom's electron be written in the form

$$V(r) = -\frac{Ze^2}{\varepsilon r} \text{ (in the Gauss system)}, \tag{21.1}$$

which is valid, generally speaking, only for macroscopic point charges. We assume the ion charge to be $+Ze$ and the electron charge $(-e)$. The total energy H of the electron is

$$H = \frac{\mathbf{p}^2}{2\mathrm{m}^*} - \frac{Ze^2}{\varepsilon r} = T + V \tag{21.2}$$

Since the values of energy obtained as a result of quantum-mechanical solution of the problem of the electron moving in a Coulomb field of a point charge coincide with those obtained from Bohr's elementary quantum theory for the hydrogen atom, we will in our quest for the energy levels of the hydrogen-like impurity atom confine ourselves to the latter.

The condition of stability of motion in a circular orbit

$$\frac{m^* v^2}{r} = \frac{Ze^2}{\varepsilon r^2} \tag{21.3}$$

enables us to find the relation between the kinetic and the potential energies

$$T = \frac{m^* v^2}{2} = \frac{Ze^2}{2\varepsilon r} = -\frac{1}{2} V, \tag{21.4}$$

whence

$$H=\frac{1}{2}V=-T=-\frac{1}{2}\frac{Ze^2}{\varepsilon r}=-\frac{p^2}{2m^*}=E. \tag{21.5}$$

Applying the condition for quantizing the Bohr momentum M

$$M=rp=rm^*v=n\hbar \tag{21.6}$$

together with the stability condition (21.3) we obtain

$$\begin{gathered}\frac{r^2m^{*2}v^2}{m^*r}=\frac{M^2}{m^*r}=\frac{n^2\hbar^2}{m^*r}=\frac{Ze^2}{\varepsilon};\\ \frac{1}{r}=\frac{m^*Ze^2}{\varepsilon\hbar^2n^2}=\frac{1}{r_n}.\end{gathered} \tag{21.7}$$

We see from here that *the radii of the Bohr orbits of a hydrogen-type atomic system in a crystal are ε times larger than those of a free "hydrogen" atom.* Allowed energy values are

$$E=E_n=-\frac{1}{2}\frac{Z^2e^4m^*}{\hbar^2n^2\varepsilon^2}. \tag{21.8}$$

The modulus of the ionization energy E_I is equal to the energy of the fundamental state ($n=1$)

$$E_I=|E_1|=\frac{1}{2}\frac{Z^2e^4m^*}{\hbar^2\varepsilon^2}. \tag{21.9}$$

The ionization energy E_I in electron-volts may be obtained by substitution of numerical values of e, $\hbar$ and m:

$$E_I=\frac{13.52\cdot Z^2}{\varepsilon^2}\left(\frac{m^*}{m}\right)\text{(eV)}, \tag{21.10}$$

where 13.52 is the energy of ionization (in eV) of a hydrogen atom. Analyze the expression (21.10) for E_I. First of all note the dependence of E_I on Z^2. It means that *the separation of the levels of singly- and double-ionized impurity atoms from the energy band is different.* This is natural since the perturbation introduced by an ion with the charge $2e$ is stronger than that caused by an ion with the charge e in full accordance with the results of generalized theory discussed in the foregoing section.

The impurity atom ionization energy decreases ε^2 times as compared to the energy of ionization of a hydrogen atom.

For germanium $\varepsilon=16$, for silicon $\varepsilon=12$, therefore in these semiconductors the impurity atom ionization energy will be, respectively, 256 and 144 times less than the ionization energy of a hydrogen atom and will comprise (in eV) a value of the order of $0.05\frac{m^*}{m}$ and $0.1\frac{m^*}{m}$. Since $\frac{m^*}{m}$ is somewhat below unity we may

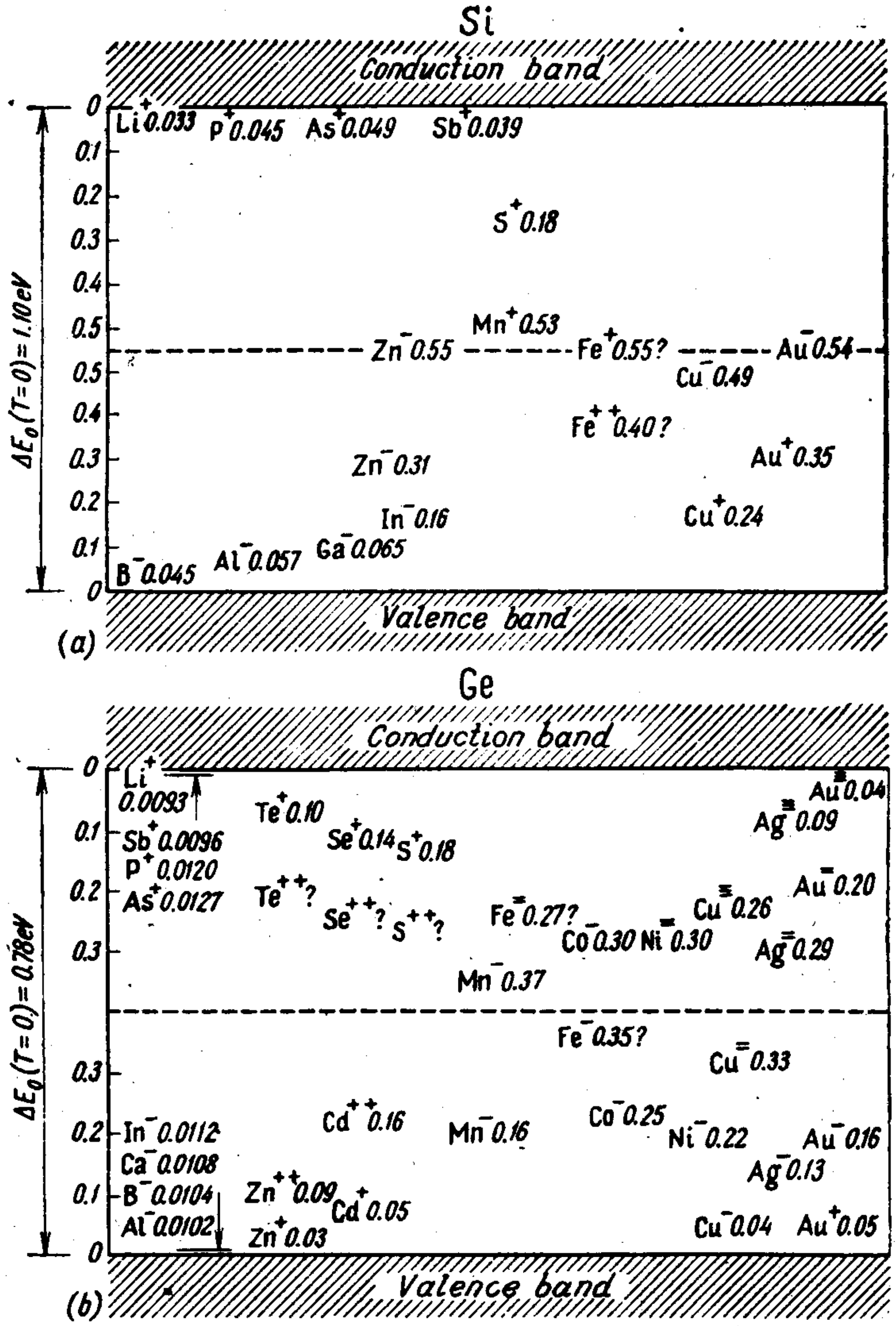


Fig. 29. A schematic diagram of impurity levels in silicon and germanium

say that in germanium impurity atom ionization energy is under 0.05 eV, and in silicon—under 0.1 eV. Figure 29 shows the levels of fundamental states of various impurities in silicon (a) and germanium (b). The difference in the ionization energy of various impurities may be assessed qualitatively by substituting into (21.10) the ionization energy of a free impurity atom instead of the ionization energy of a hydrogen atom. *The impurity atom which supplies a free electron is called a donor. It may be said that the donor level is formed below the bottom of the energy band.*

Consider the ionization energy of donors in $A^{III}B^{V}$ compounds. The dielectric permeability of these compounds is of the same order as in germanium and silicon (Table 3), so one would expect the impurity ionization energy to be of the same order, too. However, in some cases experiment proves it to be considerably less. This may be readily understood if it is taken into account that, while for germanium and silicon the ratio m^*/m is close to unity, it is, as a rule, much less in $A^{III}B^{V}$ compounds (Table 7, p. 172), and for this reason E_i may turn out to be some hundredths or thousandths fractions of an electron-volt.

Table 3

Dielectric Permeability

Material	ε	Material	ε	Material	ε
Diamond	5.7	ZnS	7.9	In As	11.7
Quartz	4.3	CdTe	11.0	Ga P	10.1
Sulphur	3.7	InSb	15.9	Ga As	12.5
Silicon	11.7	InP	10.8	Ga Sb	12.5
Germanium	15.8				

Table 4 shows E_i for impurities in $A^{III}B^{V}$ compounds and the radii a_1 of the first Bohr orbit in ångströms.

Small energy gap between the band and the donor level is a trait peculiar to some $A^{III}B^{V}$ compounds. Under normal conditions this gap is not observed experimentally. So, for example, experiments on InSb with an electron concentration $n = 10^{13}$ cm^{-3} at $T = 2$ °K did not reveal the existence of any gap between the impurity and the main bands.

Up to now we have been talking about the donor impurity which donates electrons. Consider now the acceptor impurity, for instance, indium in germanium. A neutral indium atom forms three out of four two-electron bonds. The fourth bond remains incomplete. This bond is completed by an electron passing over from one of the

Table 4

"Average" Impurity Ionization Energy E_I and the Radius of the First Bohr Orbit a_1 in $A^{III}B^{V}$ Compounds

Material	Acceptors		Donors	
	E_I, eV	a_1, Å	E_I, eV	a_1, Å
In Sb	0.03	14	0.0007	640
In As	0.05	12	0 002	310
In P	—	—	0.008	80
Ga Sb	0.03	15	0.003	150
Ga As	0.05	12	0.008	85

matrix atoms. As a result the indium atom turns into a negative indium ion.

The incomplete bond between the matrix atoms is able to migrate inside the crystal without any external energy being applied. At the same time the removal of the electron from the indium ion turns the ion into a neutral atom.

How is the periodic lattice field disturbed by the introduction of an impurity atom? Let us discuss it using a linear model as an example. Figure 22 shows the potential diagram for a chain of matrix atoms. Let us take one atom out of the chain. The potential trough in the place of this atom must disappear. But this is tantamount to a positive perturbation of the periodic lattice field. It follows that a vacancy must behave like a positive localized perturbation of sufficient magnitude embracing several neighbouring atoms and leading to the formation of a localized state close to the top of a band. In other words, *vacancies should supply holes*.

Let us, however, return to the indium impurity atom. The substitution of the vacancy by a neutral indium atom will change the slowly varying vacancy field only inside the indium atom itself; *outside the field will remain deformed with the deformation corresponding to a positive perturbation*. The separation between the top of the band and the localized level, which may be interpreted as constituting the energy of formation of a negative ion from a neutral atom, may be written in the form

$$E_I = \frac{13.52}{\varepsilon^2}\left(\frac{m^*}{m}\right), \tag{21.11}$$

where m^* is the modulus of the effective electron mass near the top of the band. Table 4 shows the values of ionization energy of acceptor impurities.

The hydrogen-type model of perturbations caused by impurity atoms is not very satisfactory, it gives only the order of impurity ionization energy and explains the appearance of discrete levels.

The above discussion shows that *point defects (and dislocations, as well) may serve as sources of carriers, usually holes.*

In addition to "shallow" levels as defined by relations (21.10) and (21.11) there are in semiconductors localized energy levels lying much further away from the energy bands. These "deep" levels cannot be explained by the hydrogen-type model. To explain the existence of deep energy levels we must presume that the electrons of impurity atoms responsible for such levels interact weakly with the matrix atoms or, put it another way, that the radius of the electron orbit of the impurity atom is small. To describe such electron states we assume for $V(r)$ a screening potential of the type

$$V(r) = -\frac{Ae^{-\alpha r}}{r}. \tag{21.12}$$

At short distances the magnitude of such a potential is great but it turns zero already at distances of the order of (2-3) α^{-1}. The greater α the narrower the potential trough that serves as a model of the impurity.

As is established in quantum mechanics the narrower the potential trough the greater the separation between the energy levels.

Assess the behaviour of an electron energy level for perturbation (21.12). To this end assess the mean perturbation energy:

$$\int_{V_R} W(r)\, d\tau = -\int_0^R \frac{4\pi Ae^{-\alpha r}}{r} r^2\, dr \approx -\frac{4\pi A}{\alpha^2}. \tag{21.13}$$

Putting potential trough radius at $\frac{c}{\alpha}$, where c is of the order of 2 to 3, we obtain for V_R

$$V_R = \frac{4\pi c^3}{3\alpha^3} \tag{21.14}$$

and for the mean perturbation W_0

$$W_0 = -\frac{3A}{c^3}\alpha. \tag{21.15}$$

Hence, *with the increase in α the diameter of the potential trough decreases, and the mean perturbation and the shift of the localized energy level both increase.* Note, by the way, that for a Coulomb potential $\alpha = 0$.

Deep levels play an important part in non-equilibrium processes.

Return now to expression (21.8) which we will re-write with reference to (21.10) in the form

$$E_n = -\frac{E_i}{n^2} \quad (n = 1,\ 2,\ 3,\ \ldots). \qquad (21.16)$$

(Energy is measured from the extremum of each band.) This means that the introduction of an impurity atom causes a whole bunch of hydrogen-like energy levels to appear inside the forbidden band. These levels converge on the corresponding band, as Fig. 30 shows for a donor impurity. The combination of impurity levels makes itself manifest in the optical properties of semiconductors (impurity light absorption, for example).

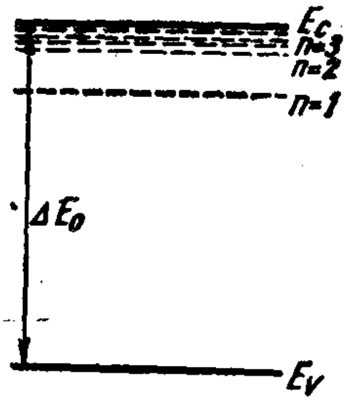


Fig. 30. A hydrogen-like impurity level system in the forbidden band

Summary of Sec. 21

1. Impurities which introduce negative perturbations into the lattice field are responsible for the appearance of discrete energy levels below the bottom of an energy band. They supply electrons and are termed donors. Impurities which introduce positive perturbations lead to the formation of discrete levels above the top of a band. They supply holes and are termed acceptors. The diagram of impurity levels is shown in Fig. 31.

2. Hydrogen-type model gives a satisfactory qualitative description of shallow impurity levels. In this case the impurity ionization energy is inversely proportional to ε^2 and directly proportional to m^*. Deep levels may be described with the aid of the screening potential.

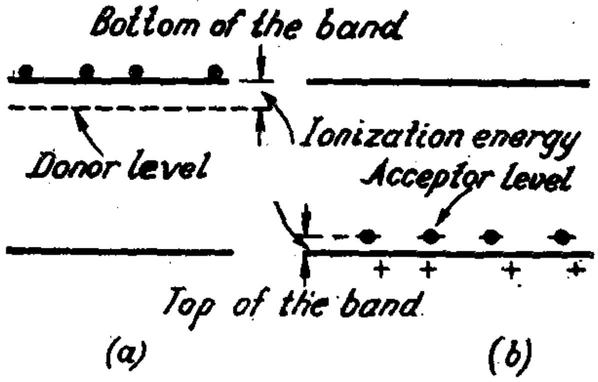


Fig. 31. Schematic diagram of donor and acceptor impurity levels

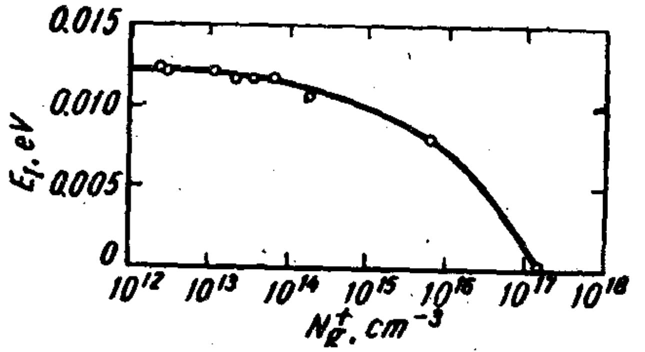


Fig. 32. The dependence of the energy of ionization of arsenic on impurity concentration in a fixed temperature interval

3. As impurity concentration increases the discrete level, because of exchange interaction, splits into the impurity band, and the impurity ionization energy decreases (Fig. 32).

22. SURFACE STATES

There is one unavoidable defect in every real crystal which stems from its finite dimensions and the presence of boundaries. Up to now we managed to exclude the boundaries by applying the condition of periodic continuation of the crystal. Let us discuss now, using the unidimensional model of the lattice field, the changes which the existence of the boundaries introduces into the energy spectrum. Place the origin of the co-ordinates at the left boundary of a unidimensional semi-infinite crystal. In the first interval ($x < 0$) $U(x) = U_0$, in the second ($x > 0$) $U(x+a) = U(x)$

The Schrödinger equations for the first and the second intervals are of the form:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi_1(x)}{dx^2} + U_0\psi_1(x) = E\psi_1(x) \quad \text{(1st interval)} \qquad (22.1)$$

$$-\frac{\hbar^2}{2m}\frac{d^2\psi_2(x)}{dx^2} + U(x)\psi_2(x) = E\psi_2(x) \quad \text{(2nd interval)} \qquad (22.2)$$

We will be interested in electron states for which $E < U_0$. Write the solution of the Schrödinger equation for the first interval:

$$\psi_1(x) = Ae^{\varkappa x} + Be^{-\varkappa x}; \quad \varkappa = \sqrt{\frac{2m(U_0 - E)}{\hbar^2}}. \qquad (22.3)$$

The second term increases infinitely with $x \to \infty$, and for this reason one should put $B = 0$. For large positive x the boundary will little affect electron motion, and the solution of (22.2) should accordingly be in the form of Bloch functions:

$$\psi_2(x) = e^{\pm ikx}\varphi_{\pm k}(x). \qquad (22.4)$$

The wave vector in the Bloch function for an unbounded crystal is a real vector, otherwise the function will be unlimited in infinity. However, since in our case the values $x < 0$ are not realized, the condition that k should be real may be violated Moreover, the conditions of continuity of the wave function and its first derivative at point $x = 0$ cannot be satisfied in case of a real k. Suppose therefore, that complexity of the wave vector provides for the satisfaction of the required conditions imposed on the wave function $\psi_2(x)$.

Accordingly, let

$$k = k_1 + ik_2 \qquad (22.5)$$

and write out the general solution for the second interval

$$\psi_2(x) = Ce^{i(k_1+ik_2)x}\varphi_k(x) + De^{-i(k_1+ik_2)x}\varphi_{-k}(x). \qquad (22.6)$$

One should choose $D=0$, otherwise $\psi_2 \longrightarrow \infty$ for $x \longrightarrow \infty$. To find A and C write out the condition of continuity of ψ and ψ':

$$\psi_1(0) = A = \psi_2(0) = C\varphi_k(0), \tag{22.7}$$

$$\psi_1'(0) = \varkappa A = \psi_2'(0) = C\left[ik\varphi_k(0) + \varphi_k'(0)\right]. \tag{22.8}$$

By dividing (22.8) by (22.7) exclude the unknown coefficients A and C to obtain:

$$\varkappa\varphi_k(0) = ik\varphi_k(0) + \varphi_k'(0), \tag{22.9}$$

or

$$\varkappa = ik + \frac{\varphi_k'(0)}{\varphi_k(0)} = ik + [\ln\varphi_k(0)]'. \tag{22.10}$$

Generally, $\varphi_k(x)$ is a complex function, and for this reason its logarithm is a complex one, too. On the other hand, $\varkappa$ is real. Separate the real and the imaginary parts of the derivative of the logarithm

$$[\ln\varphi_k(0)]' = \lambda + i\mu \tag{22.11}$$

and substitute (22.11) into (22.10):

$$\varkappa = ik + \lambda + i\mu = ik_1 - k_2 + \lambda + i\mu, \tag{22.12}$$

whence

$$\varkappa = \lambda - k_2; \quad k_1 = -\mu. \tag{22.13}$$

Write the expression for the energy:

$$E(k) = U_0 - \frac{\hbar^2\varkappa^2}{2m} = U_0 - \frac{\hbar^2(k_2-\lambda)^2}{2m}. \tag{22.14}$$

There is a definite band structure of $E(k) = E(k_1)$ for $k_2 = 0$. Possible values of $\varkappa$ in the expression for the energy are limited by the condition $\varkappa = \lambda$, where $\lambda = \mathrm{Re}[\ln\varphi_k(0)]'$, since $\varphi_k(x)|_{x=0}$ is a function of k. For $k_2 \neq 0$ and for fixed values of λ the energy turns out to be a quadratic function of k_2 for all values of λ. But this means that for $k_2 \neq 0$, *new energy values appear in the energy spectrum which were absent in the electron energy spectrum of an infinite crystal, since in the latter case they were described by the condition $k_2 = 0$. New levels due to crystal boundaries should be located in the forbidden band.*

Consider the states which correspond to these levels:

$$\psi_2(x) = Ce^{ik_1x}\varphi_k(x)\cdot e^{-k_2x} \quad (x>0); \tag{22.15}$$

$$\psi_1(x) = Ae^{\varkappa x} \quad (x<0). \tag{22.16}$$

We see from here that *the wave functions of states generated by the boundaries fall off exponentially both sides of the boundary, i.e. that these states are localized within a boundary layer whose thickness is of the order of* k_2^{-1}. Because of this such states are termed *surface states*, and the additional levels are termed *surface levels*, or *Tamm levels* after the scientist who was the first (in 1932) to forecast them. The concrete form of the expression for Tamm levels depends on the form of the potential $U(x)$ and on the position of the boundary in respect to the field through $[\ln \varphi_k(0)]'$.

For a three-dimensional crystal the number of surface states reaches the value of 10^{15} to 10^{16} cm^{-2}.

In addition to Tamm levels the surface layer contains many localized states due to numerous lattice field imperfections (defects and adsorbed atoms).

The number of states corresponding to Tamm levels is constant for a given crystal, but *the number of states due to lattice defects and adsorbed atoms may change depending on surface treatment*. Various interactions are possible between the electron states on the surface and in bulk of a crystal, and as a result *surface states greatly affect physical processes in the semiconductor*. Especially great is the effect of surface states on the operation of semiconductor devices. We shall discuss some phenomena connected with surface states in Sec. 67.

Summary of Sec. 22

1. The condition of crystal periodicity is violated because of final dimensions of a real crystal. This leads to allowed energy levels termed surface, or Tamm, levels, appearing inside the forbidden band.

2. The wave functions of Tamm levels fade exponentially on both sides of the boundary, in other words, the boundary generates localized states. The characteristic dimensions of electron localization areas are $\varkappa^{-1}$ outside and k_2^{-1} inside the crystal. Both $\varkappa$ and k_2 are related to the energy E.

3. The surface layer of a crystal contains numerous localized perturbations of the lattice introduced by crystal lattice defects and adsorbed atoms which, according to the general theory, result in the creation of localized states. They, too, are termed surface states.

4. The surface states play an important part in the functioning of semiconductor devices. The number of surface states and their nature depend on surface treatment (lapping, polishing, etching, varnishing, etc.) and on the ambient (temperature, humidity, gaseous atmosphere, etc.).

23. QUANTIZATION OF ELECTRON ENERGY IN A MAGNETIC FIELD. LANDAU LEVELS

An entire class of phenomena which take place in strong magnetic fields was excluded from the above discussion. To describe them one should discuss the Hamilton function of a particle in a magnetic field described by the vector potential $\mathbf{A}(\mathbf{r})$. The magnetic field induction $\mathbf{B}$ is related to $\mathbf{A}(\mathbf{r})$ through the expression

$$\mathbf{B} = \operatorname{rot} \mathbf{A}. \tag{23.1}$$

If

$$\mathbf{p} = m\mathbf{r} \tag{23.2}$$

is the momentum of the particle the quantity

$$\boldsymbol{P} = \mathbf{p} + e\mathbf{A} \tag{23.3}$$

will be termed *generalized momentum* of a particle with the charge e.

The Hamilton function H may be expressed in terms of the generalized momentum. To this end one should take into account that the kinetic energy

$$T = \frac{\mathbf{p}^2}{2m} = \frac{(\boldsymbol{P} - e\mathbf{A})^2}{2m} \tag{23.4}$$

and

$$H = T + U = \frac{(\boldsymbol{P} - e\mathbf{A})^2}{2m} + U(\mathbf{r}). \tag{23.5}$$

To obtain the Hamilton operator one should substitute the operator $-i\hbar\nabla$ for the momentum $\boldsymbol{P}$:

$$\hat{\boldsymbol{P}} = -i\hbar\nabla. \tag{23.6}$$

This is because for a particle moving in a magnetic (electromagnetic) field the conjugate quantities are the co-ordinate (generally, a generalized co-ordinate) and the generalized momentum. The expression (23.6) enables us to write the Hamilton operator in the form:

$$\hat{\mathbf{H}} = \frac{(-i\hbar\nabla - e\mathbf{A})^2}{2m} + U(\mathbf{r}). \tag{23.7}$$

Find the form of the operator $(i\nabla + \mathbf{A})^2$:

$$(i\nabla + \mathbf{A})^2 = (i\nabla + \mathbf{A})(i\nabla + \mathbf{A}) = -\Delta + i\nabla\mathbf{A} + i(\mathbf{A}\nabla) + \mathbf{A}^2. \tag{23.8}$$

But

$$\nabla\mathbf{A} = (\nabla\mathbf{A}) + (\mathbf{A}\nabla) = \operatorname{div}\mathbf{A} + (\mathbf{A}\nabla), \tag{23.9}$$

therefore

$$(\nabla i+\mathbf{A})^2=-\Delta+\mathbf{A}^2+2i\,(\mathbf{A}\nabla)+i\,\operatorname{div}\mathbf{A} \tag{23.10}$$

and

$$\hat{\mathbf{H}}=-\frac{\hbar^2}{2m}\Delta+\frac{e^2\mathbf{A}^2}{2m}+\frac{i\hbar e}{m}(\mathbf{A}\nabla)+\frac{i\hbar e}{2m}\operatorname{div}\mathbf{A}+U\,(\mathbf{r}). \tag{23.11}$$

Consider the free electron $[U\,(\mathbf{r})=0]$ in a uniform magnetic field $\mathbf{B}=(0,\ 0,\ B)$. Find the form of the vector potential $\mathbf{A}$:

$$\left.\begin{aligned} B_x&=\frac{\partial A_z}{\partial y}-\frac{\partial A_y}{\partial z}=0;\\ B_y&=\frac{\partial A_x}{\partial z}-\frac{\partial A_z}{\partial x}=0;\\ B_z&=\frac{\partial A_y}{\partial x}-\frac{\partial A_x}{\partial y}=B. \end{aligned}\right\} \tag{23.12}$$

The equations (23.12) will be satisfied if one puts, for example,

$$A_y=A_z=0;\quad A_x=-yB_z=-yB. \tag{23.13}$$

In this case

$$\hat{\mathbf{H}}=-\frac{\hbar^2}{2m}\,\Delta+\frac{e^2B^2}{2m}\,y^2-\frac{i\hbar eB}{m}\,y\,\frac{\partial}{\partial x}, \tag{23.14}$$

and the Schrödinger equation assumes the form

$$-\frac{\hbar^2}{2m}\,\Delta\psi-\frac{i\hbar eB}{m}\,y\,\frac{\partial\psi}{\partial x}+\frac{e^2B^2y^2}{2m}\,\psi=E\psi\,(\mathbf{r}). \tag{23.15}$$

The variables in the equation (23.15) may be separated. For $B=0$ (23.15) is actually an equation for de Broglie waves

$$-\frac{\hbar^2}{2m}\,\Delta\psi_0=E_0\psi_0, \tag{23.16}$$

with the solution

$$\psi_0\,(\mathbf{r})=\frac{1}{(2\pi)^{3/2}}\,e^{i(\mathbf{kr})};\quad E_0=\frac{\hbar^2\mathbf{k}^2}{2m}. \tag{23.17}$$

Since the dependence of (23.15) on x and z is analogous to the dependence of (23.16) we may presume that

$$\psi\,(x,\ y,\ z)=e^{i\,(k_x x+k_z z)}\cdot\varphi\,(y). \tag{23.18}$$

Substitute (23.18) into (23.15) and cancel out the exponent to obtain

$$-\frac{\hbar^2}{2m}\left(-k_x^2-k_z^2+\frac{d^2}{dy^2}\right)\varphi\,(y)+\frac{\hbar k_x eBy}{m}\,\varphi\,(y)+\frac{e^2B^2y^2}{2m}\,\varphi\,(y)=E\varphi\,(y). \tag{23.19}$$

Put

$$y = y' + y_0, \tag{23.20}$$

hence

$$y^2 = y'^2 + 2y_0y' + y_0^2$$

and the two terms of the left-hand side of (23.19) add up to

$$\frac{\hbar ek_xB}{m}y' + \frac{\hbar ek_xB}{m}y_0 + \frac{e^2B^2}{2m}y'^2 + \frac{2e^2B^2}{2m}y'y_0 +$$
$$+\frac{e^2B^2}{2m}y_0^2 = \frac{e^2B^2}{2m}y'^2 + \frac{e^2B^2}{2m}y_0^2 + \frac{eB}{m}(\hbar k_x + eBy_0)\,y' + \frac{\hbar k_xeB}{m}y_0. \tag{23.21}$$

If one chooses $y_0 = -\frac{\hbar k_x}{eB}$ the term containing y' will turn zero, and the Schrödinger equation will assume the form

$$-\frac{\hbar^2}{2m}\left(-k_x^2 - k_z^2 + \frac{d^2}{dy'^2}\right)\varphi(y') +$$
$$+\left[\frac{e^2B^2}{2m}y'^2 + \left(\frac{\hbar^2k_x^2}{2m} - \frac{\hbar^2k_x^2}{m}\right)\right]\varphi(y') = E\varphi(y'), \tag{23.22}$$

or

$$-\frac{\hbar^2}{2m}\cdot\frac{d^2\varphi(y')}{dy'^2} + \frac{e^2B^2}{2m}y'^2\varphi(y') = \left(E - \frac{\hbar^2k_z^2}{2m}\right)\varphi(y'). \tag{23.23}$$

Putting $\omega_0 = \frac{eB}{m}$ we may write

$$-\frac{\hbar^2}{2m}\cdot\frac{d^2\varphi(y')}{dy'^2} + \frac{m\omega_0^2y'^2}{2}\cdot\varphi(y') = \left(E - \frac{\hbar^2k_z^2}{2m}\right)\varphi(y') = E'\varphi(y'). \tag{23.24}$$

The equation (23.24) is an equation for a harmonic oscillator with the mass m and frequency ω_0. The energy of the oscillator is

$$E' = E'_n = \hbar\omega_0\left(n + \frac{1}{2}\right); \quad n = 0,\ 1,\ 2,\ \ldots \tag{23.25}$$

and therefore

$$E = E_n = \frac{\hbar^2k_z^2}{2m} + \hbar\omega_0\left(n + \frac{1}{2}\right). \tag{23.26}$$

The expression (23.26) shows that *the magnetic field in no way affects the motion along the magnetic field (in the direction of the z-axis)*, and the energy of this motion is not quantized. The nature of motion in *a plane perpendicular to the field—xOy—*is quite different: *the particle moves in circles with the frequency ω_0, the so-called cyclotron frequency. Energy of this motion is quantized, discrete*, the separation between the allowed values being equal to $\hbar\omega_0$. *For a quantum particle the energy of rotational motion is discrete, as distinct from a classical particle for which it is continuous.* In terms of visual definitions it may be said that all orbit radii are feasible for the classical particle while for the quantum particle the orbit radii are quantized being equal to

$$r_n \approx \sqrt{\frac{2\hbar}{m\omega_0}\left(n+\frac{1}{2}\right)} = \sqrt{\frac{2\hbar}{eB}\left(n+\frac{1}{2}\right)}. \qquad (23.27)$$

To transfer a particle from one orbit to another the energy $\hbar\omega_0 \cdot \delta n$, where δn is an integer, should be expended. Hence, the processes of energy absorption due to the changes in rotational motion should be characterized by a discrete line spectrum (this phenomenon is termed cyclotron resonance).

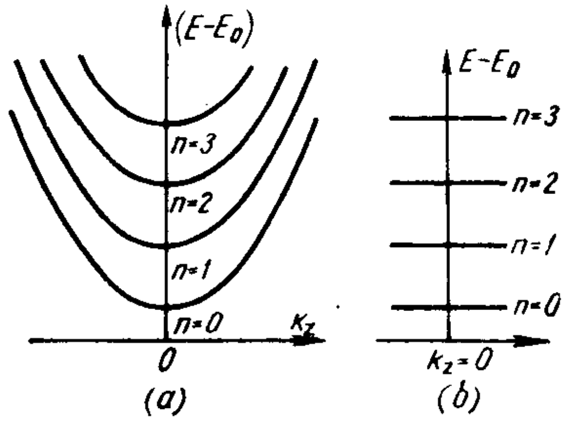


Fig. 33. The energy spectrum of the electron in a magnetic field

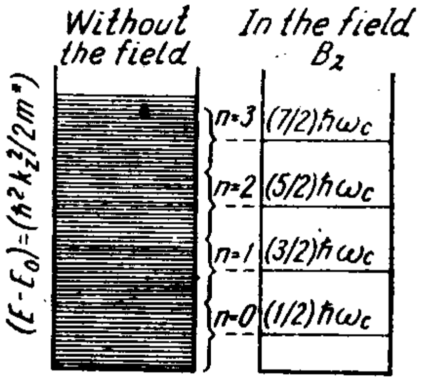


Fig. 34. The grouping of a quasicontinuous energy spectrum into discrete Landau levels

The energy spectrum $E_n(k_z)$ of a particle may often be represented in the form of a family of quadratic parabolae displaced one relative to the other along the energy axis by $\hbar\omega_0$ (Fig. 33*a*). The energy spectrum $E_n(k_z=0)$ is shown as a set of discrete levels (Fig. 33*b*). *The electron energy levels in a magnetic field are termed Landau levels.* The quantization of energy of the free electrons is the physical cause of the diamagnetism of the free electron gas.

Consider now the motion of a particle in a crystal which has been placed into a magnetic field B. The crystal Hamiltonian in this field

$$\hat{H} = \frac{(P-eA)^2}{2m} + U(r) \qquad (23.28)$$

in the effective mass approximation may be written out in the form

$$\hat{H} = \frac{(P-eA)^2}{2m^*}, \qquad (23.29)$$

which coincides with (23.4). It follows from (23.29) that *the electron energy should be quantized for each band in accordance with the expression* (23.26)

$$E=E_0+\frac{\hbar^2k_z^2}{2m^*}+\hbar\omega_0\left(n+\frac{1}{2}\right)=E(k_z,\ n,\ \omega_0). \qquad (23.30)$$

$$\omega_0=\omega_c=\frac{eB}{m^*},\ n=0,\ 1,\ 2,\ \dots$$

Figure 34 shows clearly how a quasicontinuous energy spectrum turns into a discrete one. *Energy quantization results in a whole crop of magneto-oscillatory phenomena: magneto-optical absorption, magnetic permeability oscillations, magnetoresistance oscillations, etc.* The effects of energy quantizing are noticeable when the separation between the energy levels exceeds kT, otherwise the electron distribution function remains quasicontinuous:

$$\hbar\omega_0=\frac{e\hbar B}{m^*}>kT. \qquad (23.31)$$

Substituting the values of the constants e, $\hbar$, k, m we obtain

$$1.6\cdot\frac{m}{m^*}\cdot B\cdot 10^{-20}>1.38\cdot 10^{-16}T, \qquad (23.32)$$

or

$$B>\frac{m^*}{m}\cdot T\cdot 10^4\ \text{Gs}.$$

At $T=4.2^\circ$ K (liquid helium temperature)

$$B>42\,000\left(\frac{m^*}{m}\right)\text{Gs}, \qquad (23.33)$$

i.e. $B>42\,000$ Gs for $\frac{m^*}{m}=1$; $B>4.2$ kGs for $\frac{m^*}{m}=0.1$.

Hence, to observe energy quantizing effects in magnetic fields of some tens of kilogauss one should work at liquid helium temperatures.

Summary of Sec. 23

1. Magnetic fields occasion changes in the nature of electron motion: the electron moves along the field with a velocity

$$v_z=\frac{\hbar(k_z-k_{0z})}{m^*} \qquad (23.1s)$$

and rotates in a plane perpendicular to the field.

2. The rotation frequency ω_0, termed cyclotron frequency, is determined by the magnetic field induction B and the effective mass m^*:

$$\omega_0 = \frac{eB}{m^*}. \tag{23.2s}$$

3. The energy of rotational electron motion is quantized. In the vicinity of an extremum its values are

$$E = E_0 + \frac{\hbar^2 k_z^2}{2m^*} + \hbar\omega_0\left(n + \frac{1}{2}\right) = E(k_z, n, \omega_0). \tag{23.3s}$$

4. The Hamiltonian for a free electron in a magnetic field may be derived from the expression for kinetic energy which depends on the momentum $\mathbf{p}$ if the generalized momentum

$$P = \mathbf{p} + e\mathbf{A} \tag{23.4s}$$

with the corresponding operator $-i\hbar\nabla$ is introduced.

5. The quantization of the energy of the electron in a magnetic field is the cause of numerous magneto-oscillatory phenomena.

6. A semiconductor placed into a magnetic field with induction B can absorb radiative energy with the frequency $\omega_0 = \frac{eB}{m^*}$. This phenomenon is termed cyclotron, or diamagnetic, resonance. The electron having absorbed an energy quantum $\hbar\omega_0$ goes over to a higher Landau level.

24. PAULI PRINCIPLE. CONCEPTS OF METAL, SEMICONDUCTOR AND DIELECTRIC

A set of six physical quantities, for example (x, y, z, p_x, p_y, p_z), defines the state of a classical particle. If spin is disregarded the state of a quantum particle may be described by three quantities, for instance (x, y, z), (p_x, p_y, p_z) or (E, M^2, M_z). A more accurate description with the aid of any set must include the fourth quantity—the projection of the spin on any direction, usually the z-axis, s_z. Accordingly, *the complete set needed to describe quantum states consists of four physical quantities* (L_1, L_2, L_3, s_z). The choice of the three physical quantities (L_1, L_2, L_3) is usually dictated by the type of the quantum system. For an atom, the most frequently used set consists of the energy E, the orbital angular momentum modulus $|\mathbf{M}|$ and the orbital angular momentum projection M_z. As a rule, these quantities are expressed in terms of the quantum numbers n, l and m_l. Electron states in a crystal can best be described with the aid of a set consisting of the quasimomentum projections (P_x, P_y, P_z) or of the wave vector projections (k_x, k_y, k_z) and the spin projection s_z.

At this place the reader should be reminded of how the Pauli principle is formulated: *if physical quantities* (L_1, L_2, L_3, s_z) *constituting a complete set are changed, each complete set may be obtained not more than once.* A simplified statement of the Pauli principle reads: *there may be not more than one electron in each state with a complete set* (L_1, L_2, L_3, s_z). As applied to the crystal this means that in a state (k_x, k_y, k_z, s_z) there may be not more than one electron. Since s_z may assume only two values $+1/2$ and $-1/2$ it may be said that *there may be not more than two electrons in a state* (k_x, k_y, k_z). The set (k_x, k_y, k_z) determines the value of the energy E(**k**) for each band; therefore, according to the Pauli principle, there may be not more than two electrons on any energy level.

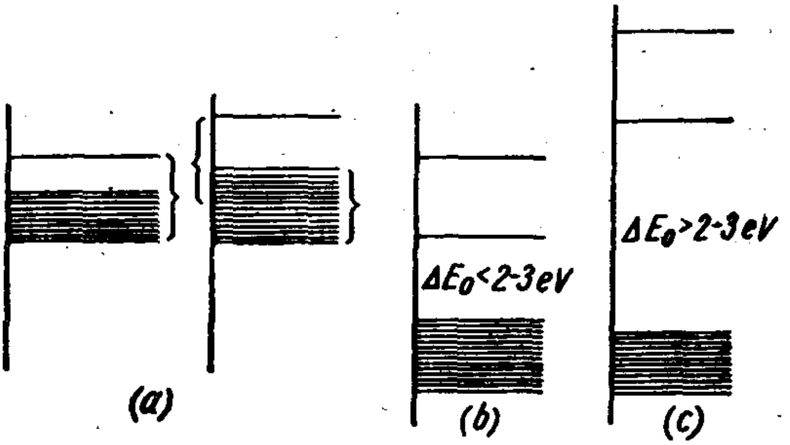


Fig. 35. Energy bands in a metal (*a*), a semiconductor (*b*) and a dielectric (*c*)

Find the number of allowed states in a band. If the crystal is of the shape of a parallelepiped with the edges N_1a_1, N_2a_2, N_3a_3, where N_i is the number of atoms in the ith edge, the total number of atoms in the crystal will be $N_1 \cdot N_2 \cdot N_3 = N$. The ith projection of the vector **k** assumes the following values:

$$k_i = \frac{2\pi}{a_i} \cdot \frac{n_i}{N_i}; \; n_i = 0; \; \pm 1; \; \pm 2; \; \ldots; \; \pm \frac{N_i - 1}{2}; \quad -\frac{N_i}{2}, \qquad (24.1)$$

where the number of different values of k_i is N_i. *The total number of different states in a Brillouin zone is* $N_1 \cdot N_2 \cdot N_3 = N$, *i.e. is equal to the number of atoms in the crystal.* This means that *a normal (non-degenerate) energy band and the corresponding Brillouin zone both contain* $2N$ *states,* and that *the maximum number of electrons in such a band is* $2N$. *If the band is degenerate with the degeneracy factor being equal to g the maximum number of electrons in it will be* $2Ng$. This means that the number of states is conserved as atoms unite to form a crystal.

Indeed, suppose the band originated from a g-fold degenerate level. The number of electrons on this level in a free atom is $2g$. If N atoms unite to form a crystal the number of states will be $2Ng$. Hence, *the number of eventual states (in a crystal) is equal to the number of initial states (in separate atoms).*

If some band contains no electrons it will play no part in conductivity when an electric field is applied. The same is true of a band all states of which are occupied by electrons. Indeed, under the action of the electric field the electron moves over to adjacent energy levels. But this is possible only if the adjacent levels are free. If, on the other hand, all the states are occupied, electron transitions from state to state will, according to the Pauli principle, become impossible, and, hence, an electric field will not be able to initiate directional motion of such an electron ensemble. *Electrons may turn into conduction electrons only in case they are in a partially occupied band. This simple conclusion serves as a basis for classifying substances by conductivity types.*

The conductors include substances which at any temperatures have partially occupied bands. The non-conductors include substances some bands of which are filled with the empty bands separated from the filled ones by an energy gap, or a forbidden band. Non-conductors may be dielectrics or semiconductors. *The division into dielectrics and semiconductors is a matter of convention. Semiconductors include substances whose forbidden band separating the highest of the filled bands (it is termed valence band) and the lowest of the empty bands (termed conduction band) does not exceed 4-5 eV.* Figure 35 shows energy band diagram for a metal (*a*), a semiconductor (*b*) and a dielectric (*c*). Consider now by way of an example the conductivity of certain substances from the point of view of their band structure.

Alkaline metals. Atoms of alkaline metals have one valence electron in the ns-state. Other electrons occupy filled shells (or subshells). To be definite, consider sodium Na as an example. Its atom has 11 electrons the states of which have the following notation: $(1s^2)\,(2s^2)\,(2p^6)\,3s$. We see that the first $1s$- and the second $2s$- and $2p$-shells are filled. The $3s$-shell has one electron. When a crystal is formed the atomic levels split into bands as is shown in Fig. 23. Ascribing to the bands the indices of the levels from which they have been formed we may say that the bands $1s$, $2s$, $2p$ are filled since they contain $2N$, $2N$ and $6N$ electrons while the numbers of states are $2N$, $2N$, $6N$, respectively. The $3s$-band, however, contains N electrons while the number of possible states is $2N$. Thus the valence band of sodium (this is true of the other alkaline metals, as well) is half occupied, the other half being free. Alkaline crystals are good conductors, i.e. metals.

Alkaline-earth elements. Alkaline-earth elements have two valence electrons in the ns-state. For example, the electron structure of Mg is $(1s^2)$ $(2s^2)$ $(2p^6)$ $(3s^2)$. If the bands of Mg were formed in the same way as of Na, Mg would have to be an insulator, since the 3s-band is filled. It is, however, a familiar fact that group two substances are good metals. This may be explained if one assumes that *the 3s- and 3p-bands overlap completely or partially, so that electrons from the upper levels of the 3s-band occupy the lower levels of the 3p-band.* The electrons in the s-band which take part in conductivity occupy the band almost completely so that some of them are in states with a negative effective mass and for this reason move in external fields as positively charged particles. *This fact manifests itself in that for many alkaline-earth metals the Hall coefficient is positive* while for electronic metals it should be negative.

Group three elements. The metallic properties of group three elements are easily explained. Since they all have three valence electrons only one of these will be in the np-state, so that when a crystal is formed the np-band will be occupied to 1/6.

Group four elements. The group four elements have four valence electrons each in $(ns^2)\,np^2$-states. If the nature of band degeneracy in the crystal would be the same as of the initial atomic levels the group four elements would have behaved as metals since the valence p-band with two electrons per atom would be occupied to 1/3. However, such elements as germanium and silicon are typical semiconductors, diamond is a dielectric and in some cases a semiconductor, and α-Sn is a semiconductor. The reason for this is that when the crystal is formed the p-band degeneracy is partially removed. The p-band splits in two—the upper band is doubly degenerate, the lower is non-degenerate (if spin degeneracy is not taken into account). The lower p-band merges with the s-band so that at definite interatomic distances the two bands—s and p—combine to form two bands, of which the upper is formed from the p-band and is doubly degenerate, and the lower is formed from both the p- and the s-bands. Thus, the upper and the lower bands each contain $4N$ states so that the lower band is filled and the upper band, empty.

The behaviour of the bands is the same in diamond, silicon, germanium and α-tin crystals except that the forbidden band width decreases in that order.

In the same way semiconducting or dielectric properties of other elementary substances may be explained.

The above examples show that *in order to explain metallic conductivity it must be assumed that the number of electrons in one of the band is less than the number of states. This may be due either to the inequality of the numbers of initial atomic states and of valence electrons, or to the overlapping of bands as a result of* *which the number of states exceeds that of valence electrons. The pre-requisite for the formation of a non-metallic crystal is a combination of bands, each band containing an equal number of states and equal number of electrons in these states.* The simplest example of the way in which this condition is achieved in the formation of A^xB^{8-x} compounds, for instance, of $A^{III}B^{V}$. The total number of valence electrons per atomic pair is eight. If the bands of the elements A and B do not overlap, the *s*- and *p*-bands of each of them (or *s*-band of one and *p*-band of the other) will contain exactly eight states per atomic pair. Consider AlP as an example. Aluminium has 13 electrons in the states $(1s^2)$ $(2s^2)$ $(2p^6)$ $(3s^2)$ $3p\,^2P_{1/2}$, phosphorus has 15 electrons distributed between the states as follows: $(1s^2)$ $(2s^2)$ $(2p^6)$ $(3s^2)$ $3p^3\,^4S_{3/2}$. $^2P_{1/2}$ and $^4S_{3/2}$ denote ground atomic states. Specifically, it may be seen from here that the sum-total momentum of unit $(l=1)$ orbital momenta of three *p*-electrons is zero and that the spins of all the three *p*-electrons are parallel $\left(1+2S_z=1+2\cdot\frac{3}{2}=4\right)$.

The ionization potentials of aluminium and phosphorus are equal to 5.99 and 11.98 V respectively, i.e. the fundamental state of the phosphorus atom lies 6 eV below that of the aluminium atom. The ionization potentials of some elements are shown in Table 5.

Since the depth of *s*-electron level in the phosphorus atom is determined by its second ionization potential we see from Table 5 that *s*-levels of aluminium and phosphorus lie considerably below the phosphorus *p*-electron levels. Therefore we may presume that *s*-bands in the AlP compound lie rather deep, and that the valence and conduction bands arise from the *p*-electron levels of aluminium and phosphorus which by partially splitting and overlapping lead

Table 5

Ionization Potentials of Some Elements

Element—Z	Valence electron state	E_1, eV	Element—Z	Valence electron state	E_1, eV
H—1	$1s$	13.595	Ga—31	$(4s^2)\ 4p$	16.11
Li—3	$2s$	5.39	Ge—32	$(4s^2)\ 4p^2$	7.88
B—5	$(2s^2)\ 2p$	8.296	As—33	$(4s^2)\ 4p^3$	9.81
C—6	$(2s^2)\ 2p^2$	11.256	Se—34	$(4s^2)\ 4p^4$	9.75
Na—11	$3s$	5.138	Cd—48	$(5s^2)$	8.991
Al—13	$(3s^2)\ 3p$	5.984	In—49	$(5s^2)\ 5p$	5.785
Si—14	$(3s^2)\ 3p^2$	8.149	Sb—51	$(5s^2)\ 5p^3$	8.639
P—15	$(3s^2)\ 3p^3$	10.484	Te—52	$(5s^2)\ 5p^4$	9.01
S—16	$(3s^2)\ 3p^4$	10.357	Hg—80	$(6s^2)$	10.43
Zn—30	$(4s^2)$	9.391	Pb—82	$(6s^2)\ 6p^2$	7.415

to the formation of triple degenerate valence and conduction bands.

It will be of interest to consider the compounds of the $A^{II}B^{VI}$-type the atoms of which have two and six valence electrons, respectively, in the following states: $(ns^2)^1S_0$ for the group II elements and $(n's^2)n'p^4\,{}^3P_2$ for the group VI elements.

The ionization potentials of these elements are very close so that for free atoms the s-levels of A^{II} and p-levels for B^{VI} lie at about the same depth. If one assumes that the valence band has been formed from the p-electron levels, the valence band should be triple degenerate and should contain 6 states per two atoms occupied by 6 electrons of the A^{II} and B^{VI} atoms. This is because the p-levels must sink lower since their wave function is less localized than that of the s-levels. The conduction band should originate from the s-levels of A^{II} atoms and should, therefore, be non-degenerate.

In the same way the nature of band occupancy of other compounds may be analyzed. For this purpose one should take into account the structure of electron shells of atoms constituting the compound. However, such an analysis enables only qualitative data about general relationships as, for example, about the relation between the width of the forbidden band and composition, to be obtained.

To explain most physical properties of semiconductors the information about the forbidden band ΔE_0, which separates the valence band from the conduction band, should be available. Table 6 contains experimental data about the forbidden band width of some semiconductors.

Summary of Sec. 24

1. The number of states in a band is equal to the number of states in atoms which constitute the crystal.

2. Each simple band may contain not more than $2N$ electrons where N is the number of atoms in the crystal.

3. Metals include substances with some partially occupied bands. Free states may be available either because the number of allowed states exceeds that of electrons, or because the filled and empty bands overlap.

4. Non-conductors include substances some bands of which are empty and the others filled. The highest filled band is termed valence band, the lowest empty band is termed conduction band. Of major importance for most physical processes is the forbidden band which separates the valence band from the conduction band. Semiconductors include substances with forbidden bands up to 4 or 5 eV wide. Substances with greater forbidden bands are dielectrics.

Table 6

Forbidden Band Widths (eV)

Type of compound	Material	Forbidden band width
Elements	Si	1.10
	Ge	0.68
	Se	2.1
	Te	0.34
	α-Sn	0.08
I-V	KSb	0.9
	K_3Sb	1.1
	CsSb	0.8
	Cs_3Bi	0.5
I-VI	Cu_2O	2.0
	Ag_2S	0.9
I-VII	CuBr	2.9
	AgI	2.8
II-IV	Mg_2Si	0.7
	Mg_2Ge	0.6
	Mg_2Sn	0.3
	Ca_2Si	1.9
	Ca_2Sn	0.9
	Ca_2Pb	0.46
II-V	Zn_3P_2	1.15
	CdSb	0.50
III-V	AlAs	2.4
	AlSb	1.5
	GaN	3.4
	GaP	2.24
	GaAs	1.4
	GaSb	0.67
	InP	1.25
	InAs	0.33
	InSb	0.18
III-VI	Al_2O_3	>5
	Al_2S_3	4.1
	Al_2Se_3	3.1
	Al_2Te_3	2.5
	GaSe	2.0
	Ga_2Se_3	1.9
	GaTe	1.5
	InSe	1.2
	In_2Se_3	1.2
IV-IV	In_2Te_3	1.0
	SiC	2.3 (cub.)
		2.9 (hex.)
II-VI	CaS	5.4
	CaSe	5.0
	CaTe	4.3
	MgSe	5.6
	MgTe	4.7
	ZnO	3.2
	ZnS	3.7
	ZnSe	2.6
	ZnTe	2.1
	SrO	5.8
	SrS	4.8
	SrSe	4.6
	SrTe	4.0
	CdS	2.4
	CdSe	1.7
	CdTe	1.5
	BaO	4.2
	BaS	4.0
	BaSe	3.7
	BaTe	3.4
	HgS ("red")	2.0
	HgSe	0
	HgTe	0
IV-VI	TiO_2	3.0
	SnO_2	4.3
	SnS	1.3
	PbO	2.3
	PbS	0.40
	PbSe	0.25
	PbTe	0.31
V-VI	As_2Se_3	1.7
	Sb_2O_3	4.2
	Sb_2S_3	1.7
	Sb_2Se_3	1.2
	Sb_2Te_3	0.3
	Bi_2O_3	3.2
	Bi_2Te_3	0.15
VI-VI	TeO_2	1.5
Other compounds	$ZnGeP_2$	2.2
	$ZnSnP_2$	2.1
	$CdGeP_2$	1.8
	$CdSnP_2$	1.5
	$ZnIn_2Se$	2.6
	$ZnIn_2Te_4$	1.4
	$CdIn_2Te_4$	0.9
	$HgIn_2Se_4$	0.6

25. MAIN FEATURES OF THE HOLE

In discussing conductivity models *we introduced the concept of hole conductivity to describe deficiency conductivity due to the absence of several of the bonding electrons. From the point of view of the band theory, hole conductivity is due to electrons of an almost filled band, i.e. to the electrons of the valence band some of the states of which are empty.* The concentration of empty states denoted by p may be assumed to represent the hole concentration. In other words, *from the point of view of statistics the hole may be defined as an empty state E (k) unoccupied by an electron.* However, such simple concepts do not determine the dynamic properties of the hole. The latter may be introduced with the aid of the Brillouin zone concept.

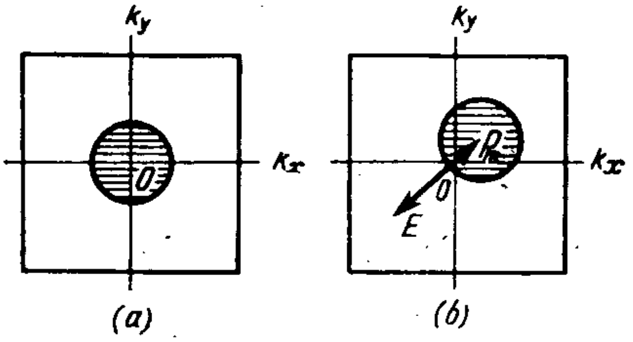


Fig. 36. The displacement of electrons in an almost free Brillouin zone under the influence of the electric field

Consider first a Brillouin zone in which only a small number of states n in the vicinity of $\mathbf{P}=0$ are occupied. Suppose, for the sake of simplicity, that a well-defined boundary separates the occupied states from the unoccupied (Fig. 36*a*). In an external field **E** all the states will be displaced by an amount $\Delta\mathbf{P}$:

$$\Delta\mathbf{P}=e^{-}\mathbf{E}\tau, \tag{25.1}$$

where τ is the mean free time.

Since the value of $\Delta\mathbf{P}$ does not depend on the initial state $\mathbf{P}_i(0)$ each occupied state is displaced within the Brillouin zone by the same amount which is equivalent to all occupied states being displaced as a whole (Fig. 36*b*):

$$\mathbf{P}_i(t)=\mathbf{P}_i(0)+e^{-}\mathbf{E}\tau \quad (t \geqslant \tau). \tag{25.2}$$

The total quasimomentum of all the electrons $\mathbf{P}_\Sigma$ is equal to

$$\mathbf{P}_\Sigma=\sum_{i=1}^{n}\mathbf{P}_i(t)=\sum_{i=1}^{n}\mathbf{P}_i(0)+ne^{-}\mathbf{E}\tau=ne^{-}\mathbf{E}\tau=n\mathbf{P}_m. \tag{25.3}$$

This is because the area of occupied states is symmetrical, so that there is a state $-\mathbf{P}_i(0)$ to correspond to each state $\mathbf{P}_i(0)$, and consequently

$$\sum_{i=1}^{n} \mathbf{P}_i(0)=0. \tag{25.4}$$

Electron velocity corresponding to state $\mathbf{P}_i$ is

$$\mathbf{v}_T(\mathbf{P}_i)=\frac{dE}{d\mathbf{P}_i}=\frac{\mathbf{P}_i}{\mathbf{m}^*} \tag{25.5}$$

and current carried by one electron

$$\mathbf{J}_i=e^-\mathbf{v}_i=\frac{e^-\mathbf{P}_i}{\mathbf{m}^*}. \tag{25.6}$$

Total current is equal to

$$\mathbf{J}=\sum_{i=1}^{n}\mathbf{J}_i=\frac{e^-}{\mathbf{m}^*}\sum_{i=1}^{n}\mathbf{P}_i=\frac{e^-}{\mathbf{m}^*}ne^-\tau\mathbf{E}=e^-n\mu_n\mathbf{E}, \tag{25.7}$$

where

$$\mu_n=\frac{e^-\tau}{\mathbf{m}^*} \tag{25.8}$$

is the electron mobility.

For a Brillouin zone corresponding to a unit crystal, n will be the electron concentration and $\mathbf{J}$, the current density, therefore $e^-n\mu_n=\sigma$ will be the specific conductivity.

Consider now the current established by carriers of an almost filled band with N' occupied and p empty states in the vicinity of $\mathbf{k}=0$ ($E(0)=E_{max}$) (Fig. 37*a*). An electric field $\mathbf{E}$ will displace all the states by an equal amount $\Delta\mathbf{P}=e^-\mathbf{E}\tau$ and therefore

$$\mathbf{P}_i(t)=\mathbf{P}_i(0)+e^-\mathbf{E}\tau. \tag{25.9}$$

The total quasimomentum will not, however, be equal to $N'e^-\mathbf{E}\tau$ as it would appear from the definition of $\mathbf{P}_\Sigma$ as a sum of $\mathbf{P}_i(t)$. This is because, owing to the displacement of every occupied state by the amount $e^-\mathbf{E}\tau$, the entire set of N' states is displaced by the same amount. The empty states are displaced by $e^-\mathbf{E}\tau$ together with the occupied ones.

As a result of the displacement of occupied states as a whole some states from the first band move over to the second band with additional empty states arising in the first band (Fig. 37*b*). All the states of the second band should be transferred to the first band by the addition of an appropriate vector $2\pi\hbar\mathbf{b}$ to $\mathbf{P}_i$. As a result the position of occupied states will be as shown in Fig. 37*c*. But in this case $\mathbf{P}_\Sigma$ will not be equal to $N'e^-\mathbf{E}\tau$. To find $\mathbf{P}_\Sigma$ one should take into account the following property of the Brillouin zones:

for a filled zone

$$\sum_{i=1}^{N} \mathbf{P}_i = 0, \qquad (25.10)$$

therefore

$$\mathbf{P}_\Sigma = \sum_{i=1}^{N'} \mathbf{P}_i = -\sum_{i=1}^{p} \mathbf{P}_i = -pe^-\mathbf{E}\tau. \qquad (25.11)$$

Thus, because of the periodicity of the states in the quasimomentum space, the total quasimomentum $\mathbf{P}_\Sigma$ of a set of N' electrons is displaced not in the direction of the force $e^-\mathbf{E}$, but in the opposite direction, i.e. in the direction $-(e^-\mathbf{E}) = e^+\mathbf{E}$. The mean quasimomentum is

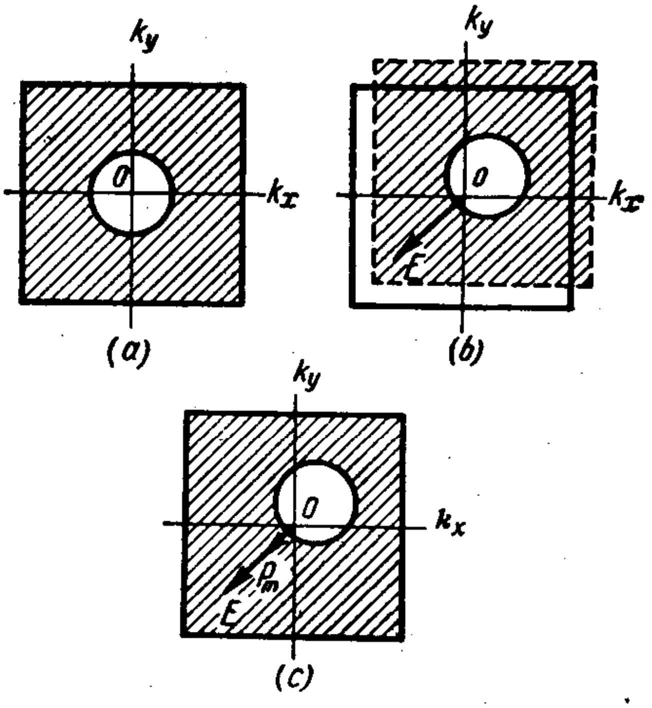


Fig. 37 The displacement of electrons in an almost filled Brillouin zone under the influence of the electric field

$$\mathbf{P}_m = \frac{\mathbf{P}_\Sigma}{N'} = -\frac{pe^-\tau}{N'}\mathbf{E}. \qquad (25.12)$$

The mean velocity is

$$\mathbf{v}_m = \frac{\mathbf{P}_m}{\mathbf{m}^*}. \qquad (25.13)$$

However, since $p \ll N' \cong N$, $\mathbf{P}_m$ will be near $\mathbf{k} = 0$, and therefore $\mathbf{m}^* < 0$. The current of N' bound electrons will be equal to

$$\mathbf{J}_b = e^- N' \mathbf{v}_m = \frac{e^- N' \mathbf{P}_m}{\mathbf{m}^*} = -e^- N' \frac{e^-\tau}{\mathbf{m}^*} \frac{p}{N'} \mathbf{E} \qquad (25.14)$$

If we define

$$-\frac{e^-\tau p}{m^* N'} = \mu_b \quad (\mu_b < 0!) \tag{25.15}$$

then

$$\mathbf{J}_b = -e^- N' \mu_b \mathbf{E}. \tag{25.16}$$

Introduce now the concept of a hole based on the following condition: *holes are quasiparticles the dynamic properties of which are identical to those of an electron ensemble*, or

$$\mathbf{P}_{\Sigma p} \equiv \mathbf{P}_\Sigma, \tag{25.17}$$

where $\mathbf{P}_\Sigma$ is the total quasimomentum of N' occupied states and $\mathbf{P}_{\Sigma_p}$—the total quasimomentum of the hole ensemble. *It is natural to define the number of holes as the number of empty states.*

It is easily seen that *the quasimomentum of a hole will in this case be defined as*

$$\mathbf{P}_p = -\mathbf{P}_0, \tag{25.18}$$

where $\mathbf{P}_0$ *is the quasimomentum of an empty state.*

Indeed, if there is one empty state in a band, then

$$\mathbf{P}_\Sigma = \sum_{i=1}^{N'} \mathbf{P}_i = \sum_{i=1}^{N'} \mathbf{P}_i + \mathbf{P}_0 - \mathbf{P}_0 = \sum_{i=1}^{N'} \mathbf{P}_i - \mathbf{P}_0 = -\mathbf{P}_0. \tag{25.19}$$

In accordance with this definition, the *transfer of quasimomentum from a hole to the lattice means the transfer of the quasimomentum of the electron ensemble to the lattice.* Now it will be easy to find the rule governing the variation of the quasimomentum: since

$$\frac{d\mathbf{P}_0}{dt} = e^- \mathbf{E}, \tag{25.20}$$

$$\frac{d\mathbf{P}_p}{dt} = -\frac{d\mathbf{P}_0}{dt} = -e^- \mathbf{E} = e^+ \mathbf{E}, \tag{25.21}$$

i.e. *the hole quasimomentum increment corresponds to a positively charged particle. Thus, the hole is a quasiparticle with $e_p > 0$.*

The expression for the current (25.14) may be re-written in the form:

$$\mathbf{J}_b = -e^- N' \frac{e^-\tau}{m^*} \cdot \frac{p}{N'} \mathbf{E} = e^+ p \frac{e^-\tau}{m^*} \mathbf{E} = e^+ p \frac{e^+\tau}{m_p^*} \mathbf{E}, \tag{25.22}$$

where

$$m_p^* = -m^* = -\frac{1}{\hbar^2} \cdot \frac{d^2E}{dk^2}. \tag{25.23}$$

In other words, *if the hole mass is defined as the electron mass with the (—) sign, the hole will have a positive mass when it is near the* *top of the valence band, and a negative mass when it is near the bottom of the band.* It follows from (25.23) that *in the case of holes* $E_p(\mathbf{k}) = -E(\mathbf{k})$, *i.e. that the dependence of energy on quasimomentum is the same as for electrons with the exception that the energy scale for holes is directed opposite to the energy scale for electrons.* This has been the cause of the change of sign of the second derivative of energy with respect to quasimomentum—*where there is an energy maximum for the electrons* ($m^* < 0$) *there is a minimum* ($m_p^* > 0$) *for holes.* Hole velocity may be defined in two ways: on the one hand

$$\mathbf{v}_p = \frac{dE_p}{d\mathbf{P}_p} = \frac{-dE}{-d\mathbf{P}} = \mathbf{v}, \tag{25.24}$$

on the other hand

$$\mathbf{v}_p = \frac{\mathbf{P}_p}{\mathbf{m}_p^*} = \frac{-\mathbf{P}_0}{-\mathbf{m}^*} = \mathbf{v}, \tag{25.25}$$

i.e., *hole velocity is the velocity of an empty state.*

We thus define the hole as a quasiparticle with the following parameters:

the charge

$$e_p = e^+ = -e^- > 0,$$

the effective mass

$$\mathbf{m}_p^{*-1} = -\mathbf{m}^{*-1} = -\frac{1}{\hbar^2} \cdot \frac{d^2E}{d\mathbf{k}^2} = \frac{1}{\hbar^2} \frac{d^2E_p}{d\mathbf{k}_p^2}; \tag{25.26}$$

the quasimomentum

$$\mathbf{P}_p = -\mathbf{P}_0; \ \mathbf{P}_{\Sigma p} = \sum_{i=1}^{p} \mathbf{P}_{ip} = \sum_{i=1}^{N'} \mathbf{P}_i = \mathbf{P}_\Sigma;$$

the velocity

$$\mathbf{v}_p = \frac{\mathbf{P}_p}{\mathbf{m}_p^*} = \frac{dE_p}{d\mathbf{P}_p} = \frac{dE}{d\mathbf{P}} = \mathbf{v};$$

the wave vector

$$\mathbf{k}_p = -\mathbf{k};$$

the energy

$$E_p(\mathbf{k}_p) = -E(-\mathbf{k});$$

the mobility

$$\mu_p = \frac{e_p\tau}{\mathbf{m}_p^*} = \frac{-e^-\tau}{-\mathbf{m}^*} = -\mu_n.$$

Thus, *all the hole parameters are determined by the parameters of the valence band electron.*

The hole drift velocity is in the direction of the field

$$\mathbf{v}_d = \mu_p \mathbf{E}. \tag{25.27}$$

At the same time the drift velocity of the bound electron ensemble is directed against the field:

$$\mu_b = -\frac{e^- \tau}{\mathrm{m}^*} \frac{p}{N'} = -\frac{e^+ \tau}{\mathrm{m}_p^*} \cdot \frac{p}{N'} = -\mu_p \frac{p}{N'}. \tag{25.28}$$

Determine the *hole distribution function* f_p:

$$f_p = 1 - f = 1 - \frac{1}{e^{\frac{E-F}{kT}} + 1} = \frac{1}{e^{\frac{F-E}{kT}} + 1}, \tag{25.29}$$

which may be written in the same form as the electron distribution function if one defines the Fermi level for holes as

$$F_p = -F. \tag{25.30}$$

Then

$$f_p = \frac{1}{\frac{E_p - F_p}{kT} + 1}. \tag{25.31}$$

In this case *the motion of the bound electron ensemble in an electric field is fully described by the motion of the hole ensemble* In the same way it may be shown that *the motion of the valence band electron ensemble in a magnetic field is described by the motion of the hole ensemble.*

Summary of Sec. 25

1. Hole conductivity is the conductivity of the electrons of the valence, or any other, band more than half the states of which are filled.

2. The hole is a quasiparticle the properties of which are determined by the properties of the valence band electron as shown by the expressions (25.26-31). The hole concentration is the concentration of empty states in the valence band.

26. BAND STRUCTURE OF SOME SEMICONDUCTORS. CALCULATION METHODS

Up to now we have been considering general methods of calculating the band pattern and have succeeded in demonstrating in a generalized form the existence of energy bands and some of their parameters. For various reasons the calculation of the band pattern of specific semiconductors is extremely complicated, first of all because there is no analytical expression for $U(\mathbf{r})$. For this reason the formulae used in all calculations contain some parameters whose values are determined by comparing theoretical and experimental data. *The forbidden width*, for example, *is determined solely from experiment*.

Nowadays, there are several methods of calculating the band pattern. We will briefly discuss some of them.

(a) **The plane wave method** (PW method). We present the periodic lattice field as a combination of the Fourier components and obtain an expansion in the form of (17.10):

$$U(\mathbf{r})=\sum_{\mathbf{b}} c_{\mathbf{b}} e^{i2\pi(\mathbf{br})}, \tag{26.1}$$

where $c_{\mathbf{b}}$ are defined by relations (17.12) and (17.9). Since $\varphi(\mathbf{r})$ is a periodic function it can be expanded into a Fourier series:

$$\varphi(\mathbf{r})=\sum_{\mathbf{g}'} a_{\mathbf{g}'} e^{i2\pi(\mathbf{g}'\mathbf{r})}. \tag{26.2}$$

The Bloch function may be represented in the form

$$\psi_{\mathbf{k}}(\mathbf{r})=\sum_{\mathbf{g}'} a_{\mathbf{g}'} e^{i(\mathbf{k}+2\pi\mathbf{g}')\mathbf{r}}. \tag{26.3}$$

Substituting the expressions for U and $\psi_{\mathbf{k}}$ into the Schrödinger equation we obtain:

$$a_{\mathbf{g}}\frac{\hbar^2(\mathbf{k}+2\pi\mathbf{g})^2}{2m}+\sum_{\mathbf{b}} c_{\mathbf{b}} a_{\mathbf{g}-\mathbf{b}}=Ea_{\mathbf{g}}. \tag{26.4}$$

The equation system (26.4) for $a_{\mathbf{g}}$ enables $E(\mathbf{k})$ to be found from the secular equation if $c_{\mathbf{b}}$ are known, i.e. if the expression for the lattice field is known. A great number of plane waves is needed to calculate $E(\mathbf{k})$, and this makes the method impractical.

(b) **The method of orthogonalized plane waves** (OPW), or the Herring method. Poor convergence of the series used to calculate $E(\mathbf{k})$ by the PW method is due to the fact that close to the ion the wave function essentially differs from a plane wave, and a great number of short wave (or large $\mathbf{k}$) plane waves is therefore needed to expand it into a series. The state close to the ions is better described by the atomic wave function $\psi_j^a(\mathbf{r}-\mathbf{n})$. A sum of such functions can be made up as follows:

$$\varphi_{\mathbf{k}j}(\mathbf{r})=\frac{1}{\sqrt{N}}\sum_{\mathbf{n}} e^{i(\mathbf{kn})}\psi_j^a(\mathbf{r}-\mathbf{n}), \tag{26.5}$$

where j is some atomic state (s, p, d, etc.), and N is the number of atoms in the crystal whose position is defined by the vector $\mathbf{n}$. Consider an expression of the form:

$$\chi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{L^{3/2}} e^{i(\mathbf{kr})} - \sum_j \mu_{\mathbf{k}j}\varphi_{\mathbf{k}j}(\mathbf{r}). \tag{26.6}$$

In the vicinity of some atom $\chi_{\mathbf{k}}(\mathbf{r})$ behaves, on the whole, as a corresponding atomic wave function. Away from the atom it is quite like a plane wave. To facilitate calculations the functions $\chi_{\mathbf{k}}(\mathbf{r})$ may be required to be orthogonal to all "fundamental" states $\varphi_{\mathbf{k}j}(\mathbf{r})$:

$$\int \varphi^{*}_{\mathbf{k}j}(\mathbf{r})\,\chi_{\mathbf{k}}(\mathbf{r})\,d\tau = 0. \tag{26.7}$$

This enables the coefficients $\mu_{\mathbf{k}j}$ to be chosen unambiguously. The OPW method makes the calculations much easier by improving the convergence of the series used to calculate $E(\mathbf{k})$.

(c) **The method of associated plane waves** (APW), or the Slater method. In the APW method the crystal wave function is chosen as a combination of plane waves "outside the atom" and of atomic wave functions $\psi^{a}_{lm}(\mathbf{r})$ "inside the atom":

$$\Phi_{\mathbf{k}}(\mathbf{r}) = \theta(r - r_i)\,a_0 e^{i(\mathbf{kr})} + \sum_{lm} a_{lm}\theta(r_i - r)\,\psi^{a}_{lm}(r). \tag{26.8}$$

The model of an atom in this case is a sphere with radius r_i; $\theta(\xi)$ is a unit step-function which sets into operation either a plane wave (for $r > r_i$), or an atomic wave function (for $r < r_i$):

$$\theta(\xi) = \begin{cases} 1 \text{ for } \xi > 0 \\ 0 \text{ for } \xi < 0. \end{cases} \tag{26.9}$$

(d) **The pseudopotential method.** Another method subsequently called the pseudopotential method was developed from the OPW method (Phillips and Kleinman) and is at present widely used for the calculation of band-patterns.

Let $\psi^{(\alpha)}_{\mathbf{k}}(\mathbf{r})$ be the desired wave function satisfying the Schrödinger equation for the crystal, and $\varphi_{\mathbf{k}j}(\mathbf{r})$, the Bloch sum made up of atomic wave functions so that it is orthogonal to the function $\psi^{(\alpha)}_{\mathbf{k}}(\mathbf{r})$:

$$\int \psi^{(\alpha)*}_{\mathbf{k}}(\mathbf{r})\,\varphi_{\mathbf{k}j}(\mathbf{r})\,d\tau = 0. \tag{26.10}$$

Define the function $\Phi(\mathbf{r})$ so that

$$\Phi(\mathbf{r}) = \psi^{(\alpha)}_{\mathbf{k}}(\mathbf{r}) - \sum_j a_{\mathbf{k}j}\varphi_{\mathbf{k}j}(\mathbf{r}). \tag{26.11}$$

Since in the vicinity of the atoms $\psi_{\mathbf{k}}^{(\alpha)}(\mathbf{r})$ behaves like one of the atomic functions which constitute the Bloch sum, $\Phi(\mathbf{r})$ will, consequently, be a smooth function everywhere, including the vicinity of the atoms. The coefficients $a_{\mathbf{k}j}$ are to be derived from the condition of orthogonality of $\psi_{\mathbf{k}}^{(\alpha)}(\mathbf{r})$ and $\varphi_{\mathbf{k}j}(\mathbf{r})$:

$$a_{\mathbf{k}j}=-\int \varphi_{\mathbf{k}j}(\mathbf{r})\,\Phi(\mathbf{r})\,d\tau=-(\varphi_{\mathbf{k}j}\Phi). \tag{26.12}$$

Find the equation for $\Phi(\mathbf{r})$. Applying the conditions

$$\hat{H}\psi_{\mathbf{k}}^{(\alpha)}(\mathbf{r})=E(\mathbf{k})\,\psi_{\mathbf{k}}^{(\alpha)}(\mathbf{r});\quad \hat{H}\varphi_{\mathbf{k}j}=E_j\varphi_{\mathbf{k}j}, \tag{26.13}$$

we obtain

$$\left\{-\frac{\hbar^2}{2m}\Delta+U(\mathbf{r})\right\}\Phi(\mathbf{r})+\sum_j a_{\mathbf{k}j}E_j\varphi_{\mathbf{k}j}=$$
$$=E(\mathbf{k})\,\Phi(\mathbf{r})+E(\mathbf{k})\sum_j a_{\mathbf{k}j}\varphi_{\mathbf{k}j}. \tag{26.14}$$

Re-write (26.14) as follows:

$$\left\{\left(-\frac{\hbar^2}{2m}\Delta+U(\mathbf{r})\right)+\sum_j\frac{a_{\mathbf{k}j}(E_j-E)\,\varphi_{\mathbf{k}j}}{\Phi(\mathbf{r})}\right\}\Phi(\mathbf{r})=E\Phi(\mathbf{r}). \tag{26.15}$$

The term $\sum_j \frac{a_{\mathbf{k}j}(E_j-E)\,\varphi_{\mathbf{k}j}}{\Phi(\mathbf{r})}=V_R(\mathbf{r})$ bears the name of repulsion potential and $V_p=U+V_R$ of pseudopotential. Now we may write:

$$\left\{-\frac{\hbar^2}{2m}\Delta+V_p\right\}\Phi(\mathbf{r})=E(\mathbf{k})\,\Phi(\mathbf{r}). \tag{26.16}$$

Since $\Phi(\mathbf{r})$ is a sufficiently smooth function it can be made up of a small number of plane waves, and they will provide for a rapid convergence of the series for calculating the dispersion dependence of $E(\mathbf{k})$. The pseudopotential may, moreover, be considered as a perturbation.

One version of the method that received the name of empiric pseudopotential method to find V_p uses experimental data, in the first instance the reflectivity spectrum and the spectrum of the imaginary part of dielectric permeability which is calculated from the reflectivity spectrum. The main advantage of the method over the others[4] is that it makes it possible to calculate with the aid of data on reflection peaks the values of $E(\mathbf{k})$ at different points of the Brillouin zone. The weakness of the method is that to obtain numerical values of interband distances it is necessary to trim the parameters (to introduce a "form factor").

(e) **The variational principle,** or the Kohn-Korringa-Rostocher (KKR) method. The quantity I as defined by the relation

$$I = \int \psi^* (\hat{H} - E)\psi\, d\tau \tag{26.17}$$

may be calculated for an arbitrary function $\psi(\mathbf{r})$, therefore I is a functional dependent on the form of $\psi(\mathbf{r})$: $I = I\{\psi(\mathbf{r})\}$. If a Hamiltonian eigenfunction is inserted into the integral I will turn zero. A small deviation of $\psi(\mathbf{r})$ from an eigenfuction will little affect the value of I. In other words, the variation of a Hamiltonian eigenfunction should provide for a zero change in the functional, i.e. $\delta I = 0$ for $\delta\psi \neq 0$.

A choice of test functions and of a potential model enables the band pattern to be calculated without parameter "trimming".

Least indeterminate are calculations which make use of the relativistic Hamiltonian; their results, however, are rather distant from those obtained with the aid of other methods. As may be seen from the above, the methods used in practice are actually combinations of the quasifree and quasibound electron methods.

(f) **k.p.-method.** The dispersion relation derived by Kane with the aid of a modified pertubation theory method is widely used to describe numerous phenomena. The essence of the latter may be understood with the aid of the relation (11.19) written in the form

$$\left\{-\frac{\hbar^2}{2m}\Delta + U(\mathbf{r}) + \frac{\hbar}{m}(\mathbf{kp})\right\}\varphi_{\mathbf{k}n}(\mathbf{r}) = \left[E_n(\mathbf{k}) - \frac{\hbar^2 k^2}{2m}\right]\varphi_{\mathbf{k}n}(\mathbf{r}). \tag{26.18}$$

The wave function and energy in the equation (26.18) have an additional index n to indicate that states belong to the nth energy band. If it can be assumed that the solution of the Schrödinger equation for some state $\mathbf{k}_0$ is known, the solution may be sought in the vicinity of that state:

$$\varphi_{n\mathbf{k}}(\mathbf{r}) = \sum_{n'} c_{n'n}(\mathbf{k} - \mathbf{k}_0)\,\varphi_{n'\mathbf{k}_0}(\mathbf{r}); \tag{26.19}$$

$$\hat{\mathbf{H}}_{\mathbf{k}_0} = \frac{\hat{\mathbf{p}}^2}{2m} + \frac{\hbar}{m}(\mathbf{k}_0\hat{\mathbf{p}}) + \frac{\hbar^2\mathbf{k}_0^2}{2m} + U(\mathbf{r}) \tag{26.20}$$

and

$$\hat{\mathbf{H}}_{\mathbf{k}_0}\varphi_{n\mathbf{k}_0} = E_n(\mathbf{k}_0)\,\varphi_{n\mathbf{k}_0}. \tag{26.21}$$

Using (26.20-21) we may re-write (26.18) in the form

$$\left\{\hat{\mathbf{H}}_{\mathbf{k}_0} + \frac{\hbar}{m}(\mathbf{k} - \mathbf{k}_0, \mathbf{p}) + \frac{\hbar^2}{2m}(\mathbf{k}^2 - \mathbf{k}_0^2)\right\}\varphi_{n\mathbf{k}} = E_n(\mathbf{k})\,\varphi_{n\mathbf{k}}. \tag{26.22}$$

Substituting (26.19) into (26.22), premultiplying by $\varphi^*_{n\mathbf{k}_0}$ and integrating over the crystal volume we obtain the equation (26.22) in matrix form

$$\sum_{n'}\left[\left\{E_n(\mathbf{k}_0)+\frac{\hbar^2}{2m}(\mathbf{k}^2-\mathbf{k}_0^2)\right\}\delta_{n'n}+\frac{\hbar}{m}(\mathbf{k}-\mathbf{k}_0\cdot\mathbf{p}_{nn})\right]c_{n'n}= \\ =E_n(\mathbf{k})\,c_{n'n}, \qquad (26.23)$$

where

$$\mathbf{p}_{n'n}=\int\varphi^*_{n'\mathbf{k}_0}(\mathbf{r})\,\hat{\mathbf{p}}\varphi_{n\mathbf{k}_0}(\mathbf{r})\,d\tau. \qquad (26.24)$$

If $\mathbf{k}$ is presumed to be close to $\mathbf{k}_0$ the term with the matrix element of the momentum $\mathbf{p}_{n'n}$ may be taken as perturbation. In this case it may be written for the second order of the usual perturbation theory

$$E_n(\mathbf{k})=E_n(\mathbf{k}_0)+\frac{\hbar}{m}(\mathbf{k}-\mathbf{k}_0,\ \mathbf{p}_{nn})+\frac{\hbar^2}{2m}(\mathbf{k}^2-\mathbf{k}_0^2)+ \\ +\frac{\hbar^2}{m^2}\sum_{n'}\frac{|\,\mathbf{k}-\mathbf{k}_0,\ \mathbf{p}_{nn}\,|^2}{E_n(\mathbf{k}_0)-E_{n'}(\mathbf{k}_0)}. \qquad (26.25)$$

If the state $\mathbf{k}_0$ corresponds to an extremum and if at the same time $\mathbf{p}_{nn}+\hbar\mathbf{k}_0=0$ the expression (26.25) will describe spheroidal constant-energy surfaces, the effective mass components along principal axes l for which are of the form

$$\frac{1}{m_l}=\frac{1}{m}+\frac{2}{m^2}\sum_{n'}\frac{({}^l p_{nn'})^2}{E_n(\mathbf{k}_0)-E_{n'}(\mathbf{k}_0)} \qquad (26.26)$$

and

$$E_n(\mathbf{k})=E_n(\mathbf{k}_0)+\frac{\hbar^2}{2}\sum_l\frac{(k_l-k_{l0})^2}{m_l}. \qquad (26.27)$$

So in this case the **k.p.**-method does lead to some results. New results are obtained when spin-orbital interaction is taken into account. The physical nature of the latter is discussed in Sec. 84, p. 623. The Hamiltonian of the spin-orbital interaction is generally of the form

$$\hat{\mathbf{H}}_{so}=\frac{\hbar^2}{4m^2c^2}([\nabla U,\ \hat{\mathbf{p}}]\,\sigma) \qquad (26.28)$$

where σ is the spin operator (Pauli matrices). In the **k.p.**-representation (26.28) should be replaced by two terms

$$\hat{\mathbf{H}}_{so}=\hat{\mathbf{H}}_1+\hat{\mathbf{H}}_2 \qquad (26.29)$$

where $\hat{\mathbf{H}}_1$ coincides with the general expression (26.28), and the second term

$$\hat{\mathbf{H}}_2=\frac{\hbar^2}{4m^2c^2}([\nabla U,\ \mathbf{k}]\,\sigma) \qquad (26.30)$$

corresponds to the linear term in (26.18) and is peculiar of the k.p.-representation. The inclusion of other terms describing spin-orbital interaction does not lead to appreciable corrections. Subsequent calculation takes account of the symmetry of $\hat{H}_0$ and $\hat{H}_{so}$ and proceeds in the assumption that the interaction is limited to that between the conduction and valence bands, the wave functions of which correspond to one s- and three p-atomic wave functions.

We will contend ourselves with citing the results of calculations of the dispersion relations. It may be obtained for the conduction band:

$$E_c(\mathbf{k}) = \frac{E_g}{2} + \frac{\hbar^2 \mathbf{k}^2}{2m} + \frac{E_g}{2}\left\{1 + 4\frac{\hbar^2 \mathbf{k}^2}{2m_c}\cdot\frac{f_1(E_c)}{E_g}\right\}^{1/2}. \qquad (26.31)$$

For the three valence bands it has been obtained:

$$E_{1v}(\mathbf{k}) = -\frac{\hbar^2\mathbf{k}^2}{2m} - \frac{\hbar^2\mathbf{k}^2}{2m_v}\left\{1-\gamma'\,\frac{k_x^2 k_y^2 + k_y^2 k_z^2 + k_z^2 k_x^2}{(k^4/3)}\right\}; \quad E_{1v}(0) = 0; \qquad (26.32)$$

$$E_{2v}(\mathbf{k}) = \frac{E_g}{2} - \frac{\hbar^2\mathbf{k}^2}{2m} - \frac{E_g}{2}\left\{1 + 4\frac{\hbar^2\mathbf{k}^2}{2m_c}\,\frac{2(E_g+\Delta)}{3E_g+2\Delta}\cdot\frac{f_2(E_{2v})}{E_g}\right\}^{1/2};$$

$$E_{2v}(0) = 0; \qquad (26.33)$$

$$E_{3v}(\mathbf{k}) = -\frac{\Delta}{2} - \frac{\hbar^2\mathbf{k}^2}{2m} - \frac{\Delta}{2}\left\{1 + 4\frac{\hbar^2\mathbf{k}^2}{2m_c}\,\frac{E_g}{3E_g+2\Delta}\,\frac{f_3(E_{3v})}{\Delta}\right\}^{1/2};$$

$$E_{3v}(0) = -\Delta \qquad (26.34)$$

The maximum energy in the valence band—that of two bands, the heavy-hole band E_{1v}.and the light-hole band E_{2v}—has been chosen as the origin of the energy scale in the expressions (26.31-34). The third valence band E_{3v} is lower by the amount Δ which is equal to the spin-orbital interaction energy. The maximum of the valence band energy and the minimum of the conduction band energy are at point $\mathbf{k}=0$, the so-called Γ (gamma) point. The Kane parameter E_g which determines the position of the bottom of the conduction band represents the forbidden band width. Negative values of E_g are permissible, too. f_1, f_2, f_3 denote slowly changing functions of the Kane parameter which assume the value equal to unity at the point Γ.

The parameter m_c is determined from the relation

$$\frac{1}{m_c} = 2\frac{\mathrm{P}^2}{\hbar^2}\,\frac{E_g + \frac{2}{3}\Delta}{(E_g+\Delta)\,E_g} \qquad (26.35)$$

where the parameter **P** is related to the matrix element of the momentum of the interacting bands

$$\mathbf{P} = -i\frac{\hbar}{m}\mathcal{P}_{cv} \qquad (26.36)$$

The parameter **P** is usually measured in eV·cm, and for many semiconductors it can be put approximately at 6×10⁻⁸ eV·cm. For "narrow-band" semiconductors where $\mathbf{k}\cdot\mathbf{P} \ll \Delta$ and $E_g \ll \Delta$ the expressions (26.31-34) assume the form:

$$E_c(\mathbf{k}) \cong \frac{E_g}{2} + \frac{\hbar^2\mathbf{k}^2}{2m} + \left\{\frac{E_g^2}{4} + \frac{2}{3}\mathbf{P}^2\mathbf{k}^2\right\}^{1/2}; \qquad (26.37)$$

$$E_{1v}(\mathbf{k}) \cong -\frac{\hbar^2\mathbf{k}^2}{2m}; \qquad (26.38)$$

$$E_{2v}(\mathbf{k}) \cong \frac{E_g}{2} - \frac{\hbar^2{}_\mathbf{k}}{2m} - \left\{\frac{E_g^2}{4} + \frac{2}{3}\mathbf{P}^2\mathbf{k}^2\right\}^{1/2}; \qquad (26.39)$$

$$E_{3v}(\mathbf{k}) \cong -\Delta - \frac{\hbar^2\mathbf{k}^2}{2m} - \frac{\mathbf{P}^2\mathbf{k}^2}{3(E_g+\Delta)}. \qquad (26.40)$$

It follows, thus, from the Kane relation that the linear term of the Hamiltonian is of some importance for narrow-band semiconductors. The dispersion law for the conduction band and for the light-hole band in case of $E_g = 0$ is linear with a three-fold energy degeneracy at point Γ. When $E_g \neq 0$ the deviation from the quadratic, or parabolic, dependence will be noticeable only for sufficiently high values of **k**.

All methods entail considerable mathematical complications. Thus, for example, the equation for calculating $E(\mathbf{k})$ at an arbitrary point **k** for germanium and silicon with the aid of the orthogonalized plane wave method is a 146th degree equation. In this case one has to make use of the group theory methods which take account of crystal symmetry. For some symmetric points the equations may be substantially simplified but still they remain equations of the 16th degree.

The information about band structure of semiconductors at present available has been obtained by theoretical, as well as by the experimental methods. Consider the band structure of some semiconductors.

Silicon. To begin with recall that a silicon atom has 14 electrons in the states $(1s^2)(2s^2)(2p^6)(3s^2)\,3p^2\,{}^3P_{012}$, i.e. two shells are filled, and the third is not. The spins of both p-electrons are parallel, and for this reason the fundamental state is a triple one. To calculate the zero approximation of the Bloch function on the basis of the quasibound electron theory *triple degenerate* (no account taken of the spins) p-state wave functions should be used. *Owing to particle interaction in the crystal the degeneracy vanishes, and* *three separate bands of E* (**k**) *dependence are established.* These bands partially overlap, and as a result of this *both the valence and the conduction bands originate from the superposition of three different bands.*

This is shown in Fig. 38 by three *E* (**k**) branches. One of the *E* (**k**) branches of the conduction band lies substantially below the others (Fig. 38*a*). *The bottom of the conduction band is determined by the position of the absolute minimum. It lies in the* [100] *direction, and for this reason the total number of equivalent energy minima is* 6. The energy minima for electrons and holes are sometimes called valleys, hence, one can say that *the conduction band of silicon has six valleys.* As is to be seen from Fig. 38*a* the dependence *E* (**k**) is different for each direction.

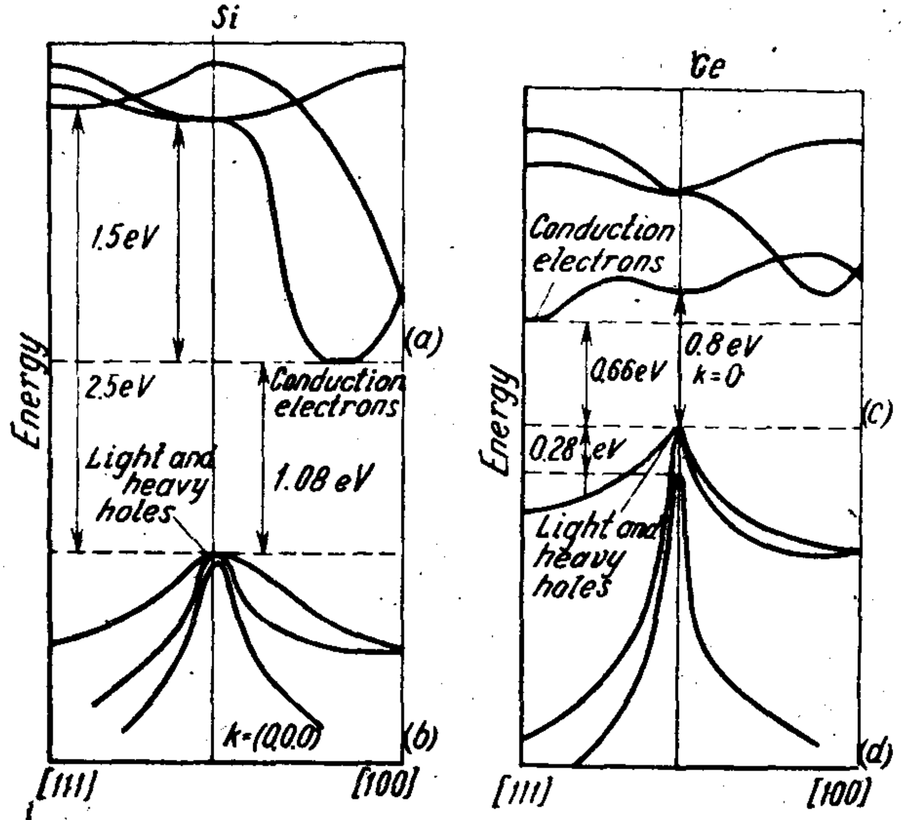


Fig. 38. The pattern of bands in germanium and silicon

Constant-energy surfaces near absolute minima are rotational ellipsoids with the axis of rotation corresponding to the [100] *direction* (Fig. 39). The dependence of energy on **k** may be represented in the form

$$E(\mathbf{k}) = E(\mathbf{k}_0) + \frac{\hbar^2[(k_1-k_{01})^2+(k_2-k_{02})^2]}{2m_1} + \frac{\hbar^2(k_3-k_{03})^2}{2m_3}, \quad (26.41)$$

$$m_1 = m_2 \neq m_3.$$

The following values for the electron effective mass tensor components for silicon have been obtained in cyclotron resonance experiments:

$$m_1 = m_2 = m_t = 0.19\ m, \quad m_3 = m_l = 0.98\ m.$$

The ratio $m_3/m_1 = 5.16$ reflects the anisotropy of constant-energy surfaces' properties. The ratio of the ellipsoids axes' lengths is

$$\sqrt{\frac{m_3}{m_1}} = 2.27.$$

The minimum point is not far away from the Brillouin zone boundary. Fig. 39 shows constant-energy surfaces for silicon near the absolute minimum.

The energy maximum for the valence band is at the centre of the Brillouin zone $\mathbf{k}_0 = 0$ *for all the three bands* with all the three bands joining at this point so that *the energy in the Brillouin zone centre becomes degenerate*. Owing to spin-orbital interaction *degeneracy partly vanishes, and one of the bands sinks by the amount* $E_{so} = 0.035$ eV. The relation between energy and the wave vector is given by the expression

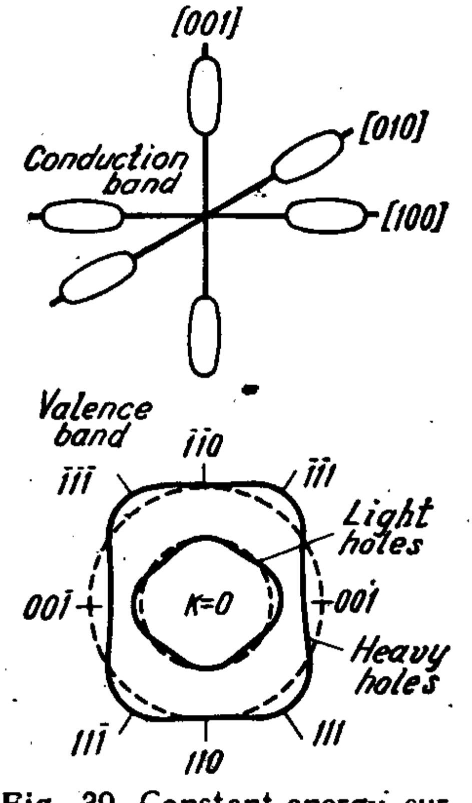


Fig. 39. Constant-energy surfaces in silicon

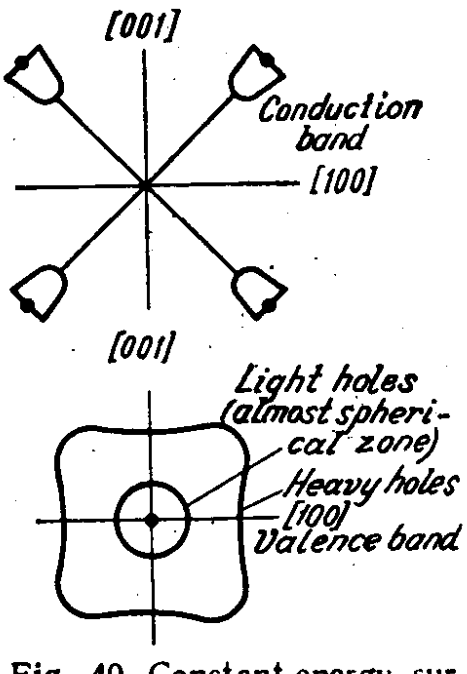


Fig. 40. Constant-energy surfaces in germanium

$$E_{1,2}(\mathbf{k}) = E(0) - \frac{\hbar^2}{2m}\left[Ak^2 \pm \sqrt{B^2k^4 + C^2(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2)}\right], \tag{26.42}$$

where A, B, C are dimensionless constants equal to 4.1 ± 0.2; 1.6 ± 0.2; 3.3 ± 0.5, respectively. Constant-enerry surfaces are deformed spheres (corrugated surfaces, see Fig. 39). The expression for $E(\mathbf{k})$ in the spherical co-ordinate system may be represented as follows:

$$E_{1,2}(\mathbf{k}) = E(0) -$$

$$-\frac{\hbar^2 k^2}{2m}\left[A \pm \sqrt{B^2 + C^2 \sin^2\theta\,(\sin^2\varphi\cdot\cos^2\varphi\cdot\sin^2\theta + \cos^2\theta)}\right]. \qquad (26.43)$$

For a given $\mathbf{k}$ the expression under the root sign may obviously vary from B^2 to $B^2 + \frac{5C^2}{16}$ with corresponding changes in energy from $A \pm B$ to $A \pm \sqrt{B^2 + \frac{5C^2}{16}}$ $\left(\text{in } \frac{\hbar^2 k^2}{2m} \text{ units}\right)$; i.e., from 5.7 to 6.5 and from 2.5 to 2.7, respectively. The second derivative at point $\mathbf{k} = 0$ is equal to $\frac{1}{\hbar^2}\frac{d^2E}{d\mathbf{k}^2} = \frac{A \pm B}{m}$ which is equivalent to one obtained when the terms with C are neglected, and this is not permissible. Averaging the expression for $E_{1,2}$ over the angles we may write

$$E_{1,2} = E(0) - \frac{\hbar^2 k^2}{2m}\left[A \pm \sqrt{B^2 + \frac{C^2}{5}}\right],$$

i.e. the corrugated surface is supplanted by some "mean" geometrical surface. In this case we assume the effective mass to be a scalar:

$$m_p^* = \frac{m}{A \pm \sqrt{B^2 + \frac{C^2}{5}}}, \qquad (26.44)$$

and there should be as a result *two types of holes*: the *heavy*

$$m_h^* = \frac{m}{A - \sqrt{B^2 + \frac{C^2}{5}}}; \quad m_h^* = 0.52m; \qquad (26.45)$$

and the *light*

$$m_l^* = \frac{m}{A + \sqrt{B^2 + \frac{C^2}{5}}}; \quad m_l^* = 0.16\,m \qquad (26.46)$$

with the ratio between their masses being

$$\frac{m_h^*}{m_l^*} = 3.3. \qquad (26.47)$$

Experimental values are $m_h^* = 0.49m$, $m_l^* = 0.16m$ and $m_h^*/m_l^* = 3.1$.

The expression for the third branch is

$$E_3(\mathbf{k}) = E(0) - E_{so} - \frac{\hbar^2 k^2}{2m} A; \quad E_{so} = 0.035 \text{ eV}. \qquad (26.48)$$

The effective mass of the holes of the third type m_s^* is a scalar, it is equal to $m_s^* = \frac{m}{A}$; $m_s^* = 0.24m$, i.e. to the mean of the light and heavy masses. However, holes with this mean mass have not been observed in experiments since the corresponding energy band is 0.035 eV below the band for the light and heavy holes.

The minimum distance between the bottom of the conduction band and the top of the valence band corresponds to the distances between different points of the Brillouin zone (Fig. 39). *It is this minimum distance that bears the name of forbidden band and determines the course of all thermal excitation processes.*

Germanium. The conduction band of germanium consists of three overlapping energy bands. *The absolute minimum is at the points of Brillouin zone which correspond to the* [111] *direction. There are, accordingly, eight identical minima.* The co-ordinates of the minima are ($^1/_2$, $^1/_2$, $^1/_2$) in $\frac{2\pi}{a}$ units, i.e. *they belong to the Brillouin zone boundary. Constant-energy surfaces are rotational ellipsoids with the axis of rotation corresponding to the* [111] *crystallographic direction* (Fig. 40).

The expression for energy is of the form similar to that for silicon:

$$E(\mathbf{k}) = E(\mathbf{k}_0) + \frac{\hbar^2 (k_1 - k_{01})^2 + \hbar^2 (k_2 - k_{02})^2}{2m_1} + \frac{\hbar^2 (k_3 - k_{03})^2}{2m_3}. \qquad (26.49)$$

Cyclotron resonance experiments give

$$m_1 = m_2 = m_t = (0.082 \pm 0.001)\, m; \quad m_3 = m_l = (1.58 \pm 0.04)\, m.$$

The ratio of the squares of ellipsoids axes is

$$\frac{m_3}{m_1} = \frac{m_l}{m_t} = 19.3; \text{ and } \sqrt{\frac{m_3}{m_1}} = 4.4.$$

The dependence of energy on **k** for germanium is shown in Fig. 38*c* for two different directions in the Brillouin zone. Since the minimum belongs to the zone boundary the *first Brillouin zone can claim only half of the energy ellipsoid, and, as a result, germanium has only four energy ellipsoids instead of eight.*

The structure of the valence band in germanium is quite identical with that in silicon (Fig. 38*b*; 40). The averaged effective masses of light and heavy holes may be derived from the constants: $A = 13.0 \pm 0.2$; $B = 8.9 \pm 0.1$; $C = 10.3 \pm 0.2$. From experiment $m_l^* = 0.04m$; $m_h^* = 0.34m$. *The third band drops by the amount of* *spin-orbital interaction* $E_{so} = 0.28$ eV. Holes of medium mass are not observed experimentally.

Intermetallic compounds. The compounds of group III and group V elements ($A^{III}B^{V}$) termed intermetallic are at present growing in importance because of some valuable properties. Best known among them are GaAs; InSb; GaP and GaSb.

Energy band structure of intermetallic compounds is calculated by comparing it to the structure of group IV element crystals. This is because the lattice of intermetallic compounds is of the zinc-blende type as compared to the diamond type for group IV element crystals. It is a known fact that the position of the sites of both lattice types in space is quite similar except that in the zinc-blende type the positions of A and B atoms alternate while the diamond type lattice consists of identical atoms. For this reason the periodic lattice field of an $A^{III}B^{V}$ compound has no inversion centre, i.e. $U(\mathbf{r}) \neq U(-\mathbf{r})$. It is assumed that the $A^{III}B^{V}$ lattice field may be represented as a combination of group IV lattice fields and an antisymmetrical term which may be regarded as a perturbation leading to changes in the known band pattern of group IV element crystals. Denoting the potential energy of $A^{III}B^{V}$ compounds by $U^{III-V}(\mathbf{r})$ and the potential energy of $A^{IV}B^{IV}$ "compounds" by $U^{IV-IV}(\mathbf{r})$ we obtain

$$U^{III-V}(\mathbf{r}) = U_s^{IV-IV}(\mathbf{r}) + U_a^{IV-IV}(\mathbf{r}) + [\Delta U_s(\mathbf{r}) + \Delta U_a(\mathbf{r})], \qquad (26.50)$$

where

$$U_s^{IV-IV}(\mathbf{r}) = U_s^{IV-IV}(-\mathbf{r}); \quad U_a^{IV-IV}(\mathbf{r}) = -U_a^{IV-IV}(-\mathbf{r}),$$

i.e. the field of $A^{IV}B^{IV}$ "compounds" is divided into a symmetrical U_s and an antisymmetrical U_a parts. For an elementary substance $U_a(\mathbf{r}) = 0$, and the symmetrical and antisymmetrical parts of the perturbation are ΔU_s and ΔU_a, respectively. For the elements of the same period of the Mendeleyev table the perturbation should, generally, be antisymmetrical, and for this reason the correction to the $A^{IV}B^{IV}$ "compound" energy should, to the first approximation, be negligible, and, accordingly, second approximation should be used. The calculation for BN is based on the diamond (C) structure, for GaAs on Ge, for AlP on Si, and for InSb on α-Sn.

An analysis of available experimental data together with theoretical calculations enables the following conclusions as to the energy band pattern of intermetallic compounds to be drawn. *Constant-energy surfaces both for the valence and conduction bands are spheres by force of which the effective masses both of electrons and holes are scalars. The bands are made up of three different bands since they originate from the p-levels.*

There is some distinction between the valence band of intermetallic compounds and that of group IV crystals—the former, too, consists of three bands one of which, owing to spin-orbital interaction, is below the others by the amount E_{so} which is different for different compounds. *Two remaining bands corresponding to light and heavy holes split in the centre of the Brillouin zone because of the presence of an antisymmetrical part of the lattice field. This leads to a displacement of energy maxima of light and heavy holes relative to the Brillouin zone centre. Such a displacement may take place for the maxima of one, or both bands.* However, the displacement is, as a rule, very small, so that $E(0)$ is below the maximum energy by some hundredths or even thousandths of an electron-volt. InSb and InAs have the minimum displacement (possibly, equal to zero). The displacement of valence band maxima in relation to the Brillouin zone centre is usually in the [111] or [100] direction (Fig. 41). The absolute minimum may be both in the centre of the Brillouin zone and at other points (Fig. 42). Conduction band contains three E (k) bands. *Because of small electron effective mass energy increases rapidly with the increase of the wave vector* and already at relatively small values of k is determined not only by the quadratic term $\frac{\hbar^2 k^2}{2m^*}$ but also by terms of higher powers of k. As a result electron effective mass turns out to be dependent on electron concentration. If some "generalized" effective mass is defined by the relation

$$E(k) = E(0) + \frac{\hbar^2 k^2}{2m^*(\mathbf{k})}, \qquad (26.51)$$

the effective mass may be said to depend on **k**.

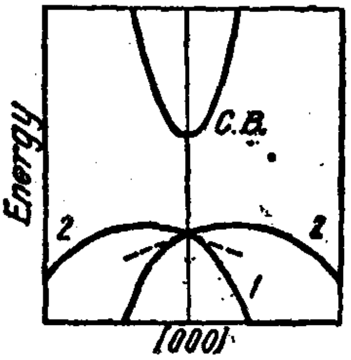


Fig. 41. The band pattern of some $A^{III}B^{V}$-type semiconductors

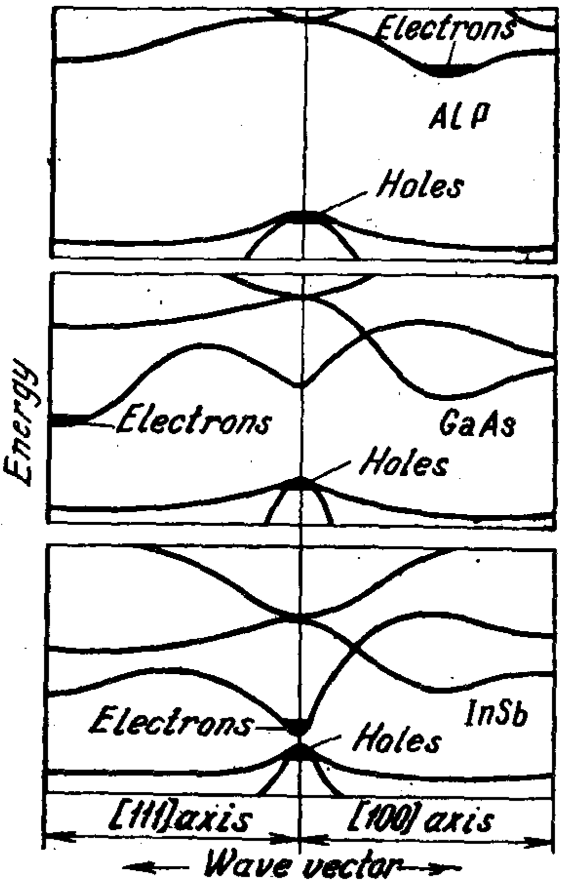


Fig. 42. The band pattern of $A^{III}B^{V}$ compounds

Table 7

Effective Masses of Electrons $\frac{m_n^*}{m}$ and holes $\frac{m_p^*}{m}$

Material	m_n^*/m	m_p^*/m	Material	m_n^*/m	m_p^*/m
Germanium	$m_l = 1.3$ $m_t = 0.082$ for $k = 0$:0.036	0.34 0.04 0.07	Gallium arsenide	0.07	0.5
			Gallium antimonide	0.05	0.5
Silicon	$m_l = 0.97$ $m_t = 0.19$	0.5 0.16 0.25	Indium antimonide	0.013	0.6
			Indium arsenide	0.03	0.4
			Indium phosphide	0.07	0.4

Table 7 shows the values of effective masses of electrons and holes in germanium, silicon and $A^{III}B^{V}$ compounds.

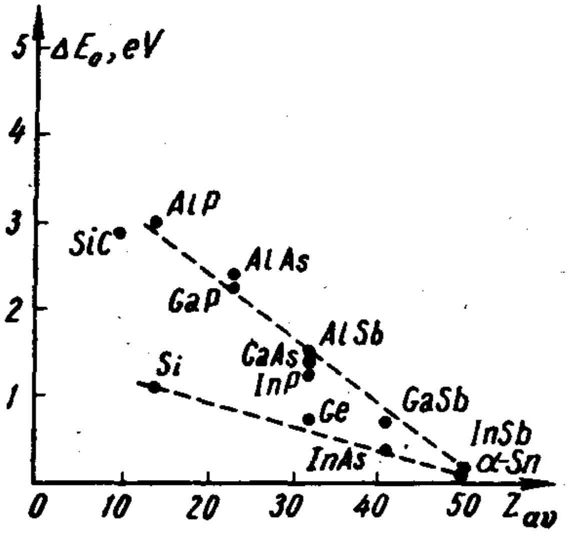
Many of the finer details of the band pattern and of the carrier parameters still await clarification.

In conclusion we are about to make some remarks concerning the forbidden band width.

The forbidden band width in elementary group IV crystals decreases with the increase of the charge of the nucleus. In $A^{III}B^{V}$ compounds, too, the forbidden band width decreases with the mean nuclear charge as is shown in Fig. 43*a*. In addition the figure shows that the forbidden band of $A^{III}B^{V}$ compounds whose mean nuclear charge is equal to the nuclear charge of group IV element is approximately twice as wide as the forbidden band of the corresponding group IV element crystals as, for instance, is the case of ${}_{13}Al_{15}P$ with the mean charge of 14 and of ${}_{14}Si$ with the same charge.

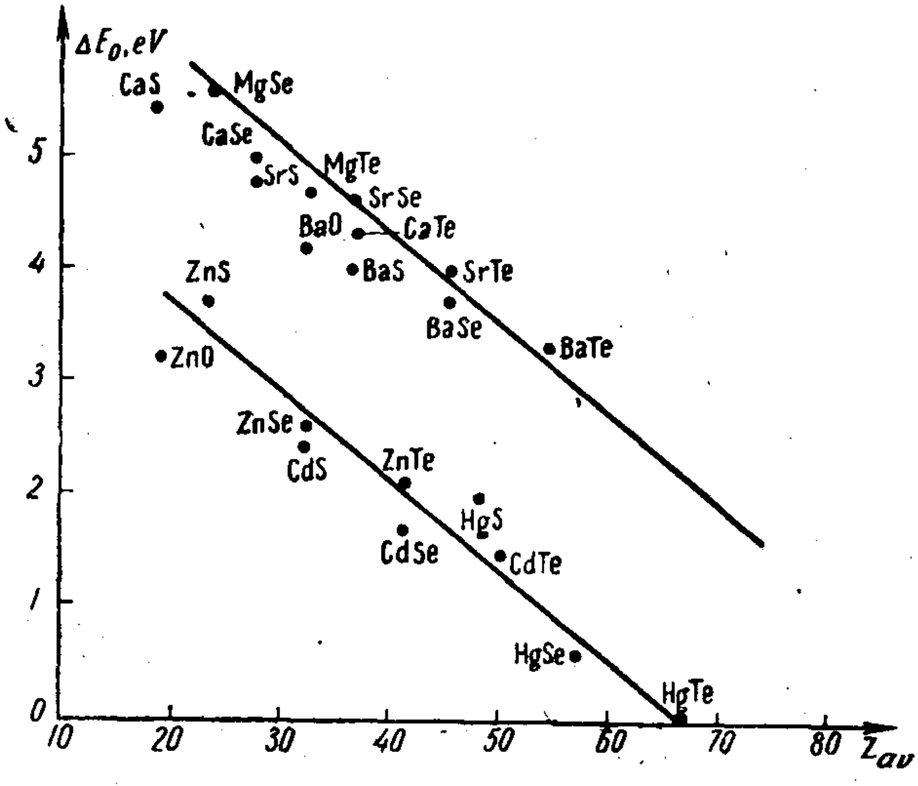
The former fact is easily explained both in terms of the quasibound and the free electron theories. In the quasibound electron theory this results from the increase in the value of the exchange integral since the dimensions of the electron cloud (the electron shell) are greater for higher *Z*. In the free electron theory the explanation may be based on the more rapid change of potential energy for smaller nuclear charges and, correspondingly, for smaller electron clouds.

The explanation of greater forbidden band widths in $A^{III}B^{V}$ compounds as compared to that of $A^{IV}B^{IV}$ compounds of equal mean



(a)

Fig. 43a. The relation of the forbidden band width with the average nuclear charge of the $A^{III}B^{V}$ compounds



(b)

Fig. 43b. The relation of the forbidden band width with the average nuclear charge in $A^{II}B^{VI}$ compounds

nuclear charge is that the potential energy changes more rapidly in the interatomic space of $A^{III}B^{V}$ compounds because the chemical bond there is partly of ionic nature. Since the charge of valence electrons is more evenly distributed in the interatomic space of $A^{IV}B^{IV}$ "compounds" than is the case with $A^{III}B^{V}$ compounds, the co-ordinate dependence of potential energy in the former will be smoother, and this will result in smaller forbidden band width.

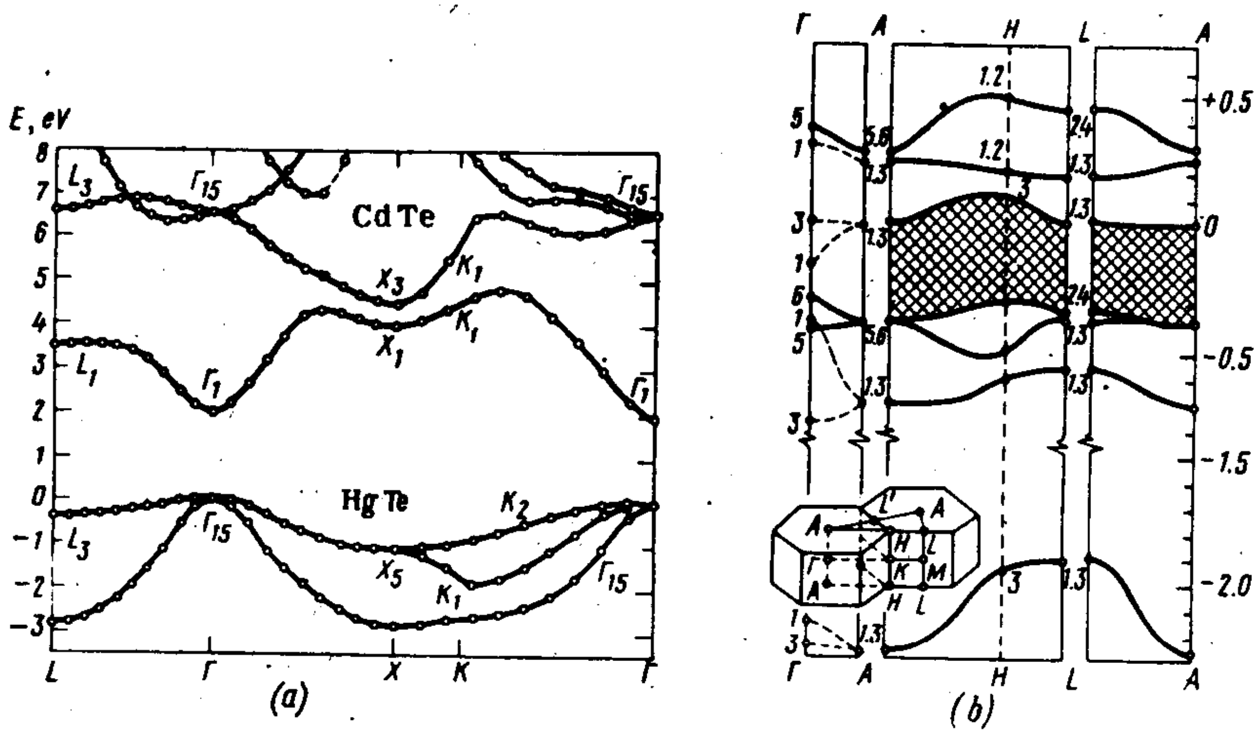


Fig. 44. Energy band pattern of some $A^{II}B^{VI}$ compounds: cadmium and quicksilver tellurides (*a*) and cadmium selenide (*b*) calculated with the pseudopotential and the OPW methods (solid line—pseudopotential, circles—OPW)

Consider the relation between the forbidden band width and the mean charge in $A^{II}B^{VI}$ compounds. It may be seen from Fig. 43*b* that the values of ΔE_0 for each subgroups of group II elements fit a straight line well.

The graphs are represented by parallel straight lines displaced relative to each other by approximately 2.2 eV along the ΔE_0-axis. This amount is related to the ionization energy of free atoms of subgroup I (A_I^{II}) and of subgroup II (A_{II}^{II}) of group II. The mean ionization potential for A_{II}^{II} atoms is 9.6 eV which is comparable to the mean ionization potential of B^{VI} elements (subgroup II) of 10.5 eV, the corresponding value for A_I^{II} being only 5.9 eV.

If it is assumed that the valence band of $A^{II}B^{VI}$ is formed from the fundamental state of B^{VI} and the conduction band—from that of A^{II}, it will be easily understood that the forbidden band in $A_I^{II}B^{VI}$ will have to be wider than in $A_{II}^{II}B^{VI}$ because the fundamental level in A_{II}^{II} is above that of B^{VI} by 1 eV on the average, the same value for A_I^{II} being equal to 4.6 eV. Generally speaking, it may be expected that the conduction band in $A_I^{II}B^{VI}$ may originate from an excited state level of B^{VI} while in $A_{II}^{II}B^{VI}$ its origin is apparently the fundamental state of A_{II}^{II} atoms.

Figure 44 shows the pattern of energy bands of cadmium and quicksilver tellurides calculated by different methods. The band extrema are at the point Γ at which the valence band is doubly degenerate. The position of the third valence band is determined by spin-orbital splitting (Fig. 44*a*). In the wurzite lattice this degeneracy is relieved by the hexagonal crystal (Fig. 44*b*).

27. QUASIPARTICLE CONCEPT

The origin of the band pattern in the previous discussion was traced to the periodic nature of the lattice field, i.e. to *long-distance order* in the arrangement of atoms. However, *in semiconductors the band pattern of energy is also observed when long-distance order is disturbed*, for instance, in the molten state. Hence, *long-distance order, i.e. periodic nature of potential energy in solids, is not a prerequisite for band formation — this is only one of sufficient conditions which became essential only because of the single-electron approximation*. Apparently, an appropriate field pattern at short distances (short-distance order) is already a sufficient condition, but the proof of this involves the solution of the multiparticle problem.

To demonstrate that long-distance order is not necessary for energy bands to be formed consider the following problem again. Suppose we have a system of N atoms at varying distances from each other. Suppose also that all distances are sufficiently large for atomic interaction to be neglected. In this case there is an obvious solution for a system of N atoms — the wave function of the system is a product of the atomic functions, and the energy of the system is the sum of the energies of atomic electrons. Such a state of the atomic system is degenerate since an interchange of electron states of different atoms is possible without a change in the energy. The degeneracy is $N!$-fold.

Draw the atoms together so that they interact. Considering the interaction of atomic pairs we will obtain $(N-1)$ different functions $W(\mathbf{r}-\mathbf{n})$ providing additional potential energy for the nth atom. It is quite obvious that main part in the value of $W(\mathbf{r}-\mathbf{n})$ will be played only by the nearest neighbours of the nth atom. The set of the $W(\mathbf{r}-\mathbf{n})$ functions will be the perturbation for the degenerate system.

It is, however, an established fact that a perturbation imposed on a degenerate system removes degeneracy totally or partly. This means that now the *energy of electrons belonging to different atoms* *will be different depending on the form of $W(\mathbf{r}-\mathbf{n})$ the latter being determined solely by short-distance order. In case degeneracy vanishes completely each level splits into N sublevels. For large N the sublevels, practically, merge into a quasicontinuous energy band.* If some $W(\mathbf{r}-\mathbf{n})$ turn out to be the same the corresponding solution will remain degenerate, but, generally, with a new (as compared to the initial) energy value.

The degeneracy factor of some sublevel will be equal to the number of identical $W(\mathbf{r}-\mathbf{n})$ functions. Introducing some parameter $\mathbf{k}$ to define the distribution of $W(\mathbf{r}-\mathbf{n})$ we obtain the dependence of energy on $\mathbf{k}$. The quantity $\mathbf{k}$ may be connected with the motion of electrons inside an atomic system via an exchange mechanism. Such qualitative considerations make it understandable that any system of interacting atoms may have a band energy pattern. *It is this consideration that enables the impurity band concept to be introduced since the distribution of impurity atoms over the crystal is non-uniform.*

The description of the properties of a solid as a system of particles is considerably facilitated by the introduction of quasiparticles which are the bearers of properties reflecting the properties of the solid. *Electrons and holes the properties of which have been described above are just such particles.*

There may be quasiparticles of other types in a solid besides the charge carriers. *The uncharged (non-current) quasiparticles include excitons and polarons.*

The transition of an electron from the valence to the conduction band is tantamount to the "ionization" of a matrix atom. To this end the atom must receive energy not less than the width of the forbidden band. If an atom is able to absorb energy less than the forbidden band width it will retain its electrons and itself go over into an excited state. There must be an energy level inside the forbidden band to correspond to this atomic (to be precise, crystal) state. Frenkel termed this excited state of an atom an exciton. *Excitation energy may be transmitted from atom to atom, and in this way an exciton may travel along the crystal. The motion of an exciton means the transport of atomic excitation inside the crystal, but not of the excited atom itself.* If several excited states are possible the forbidden band should contain several different energy levels. Exciton levels are similar to impurity levels.

The definition of the exciton may be approached from another side. Suppose the energy of an electron transferred from the valence to the conduction band is E_n, and the energy of the hole it left behind E_p. There must be an attraction between the hole and the electron the energy of which is

$$V(\rho) = -\frac{e^2}{\varepsilon|\mathbf{r}_n-\mathbf{r}_p|}; \ \rho=|\mathbf{r}_n-\mathbf{r}_p| \text{ (in the Gauss system)}, \quad (27.1)$$

where $\mathbf{r}_p$ is the hole co-ordinate and $\mathbf{r}_n$—the electron co-ordinate. The motion of a system of two bodies may be reduced to the motion of one particle in a force field:

$$V(\rho) = -\frac{e^2}{\varepsilon\rho}; \quad m_{np}^* = \frac{m_n^* m_p^*}{m_n^* + m_p^*}. \tag{27.2}$$

This particle will have so-called reduced mass m_{np}^* the motion of which will be described in the co-ordinate system linked with the inertia centre by the equation:

$$\left(-\frac{\hbar^2}{2m_{np}^*}\Delta - \frac{e^2}{\varepsilon\rho}\right)\psi(\xi, \eta, \zeta) = \tilde{E}\psi(\xi, \eta, \zeta) = E^{ex}\psi(\xi, \eta, \zeta), \tag{27.3}$$

where

$$\Delta = \frac{\partial^2}{\partial\xi^2} + \frac{\partial^2}{\partial\eta^2} + \frac{\partial^2}{\partial\zeta^2}; \quad \rho = (\xi^2 + \eta^2 + \zeta^2)^{1/2}. \tag{27.4}$$

The equation (27.3) is identical to the equation for a hydrogen-like system. The expression for $\tilde{E}$ may be written in the same way as for the hydrogen atom. The energy for a free hydrogen atom is of the form:

$$E_n^{\mathrm{H}} = -\frac{me^4}{2\hbar^2 n^2} = -\frac{E_{\mathrm{I}}^{\mathrm{H}}}{n^2}, \tag{27.5}$$

where E_{I} is the ionization energy of a hydrogen atom in the fundamental state ($n = 1$). The origin of the energy scale is assumed to be the energy of a stationary electron at an infinite distance from the atom.

If some other value, for instance, the energy of the fundamental state, is taken as the origin, i. e. $E_{\mathrm{I}}^{\mathrm{H}} = 0$, then

$$E_n = E_{\mathrm{I}}^{\mathrm{H}} - \frac{E_{\mathrm{I}}^{\mathrm{H}}}{n^2} = E_{\mathrm{I}}^{\mathrm{H}}\left(1 - \frac{1}{n^2}\right). \tag{27.6}$$

The energy E_n^{ex} of a free electron-hole pair may be written in the same form as for (27.5):

$$E_n^{ex} = -\frac{E_{\mathrm{I}}^{ex}}{n^2}; \quad E_{\mathrm{I}}^{ex} = \frac{m_{np}^* e^4}{2\hbar^2 \varepsilon^2}. \tag{27.7}$$

To determine the position of the exciton levels in the energy scale of the crystal as a whole, it has to be taken into account that for $n = \infty$ and $\rho = \infty$ $E_\infty^{ex} = E_v + E_c$ and not zero. Therefore

$$E_n^{ex} = E_v + E_c - \frac{E_{\mathrm{I}}^{ex}}{n^2}. \tag{27.8}$$

If the energy is measured from the top of the valence band: $E_v = 0$;

$E_c = \Delta E_0$, then $E_n^{ex} = \Delta E_0 - \frac{E_I^{ex}}{n^2}$. The fundamental state is $E_1^{ex} = \Delta E_0 - E_I^{ex}$, which may be assessed as follows:

$$E_I = \frac{m_{np}^* e^4}{2\hbar^2 \varepsilon^2} = \frac{13.5}{\varepsilon^2} \cdot \left(\frac{m_{np}^*}{m}\right) \text{(eV)}. \tag{27.9}$$

If $m_n^* = m_p^* = m$, then $m_{np}^* = \frac{m}{2}$, and we see that the exciton levels are near the bottom of the conduction band. The Bohr orbit radius for the exciton is

$$a_n^{ex} = \frac{n^2 \hbar^2}{e^2 m_{np}^*} \varepsilon = 0.528 \cdot \left(\frac{m}{m_{np}^*}\right) \varepsilon \text{ (Å)}. \tag{27.10}$$

The smaller the effective mass and the greater the dielectric permeability, the greater the exciton radius and the smaller the ionization energy, and for that reason it is not easy to observe excitons in substances of great ε and small m_n^*, m_p^*. Large radius excitons are known as Mott excitons.

For a bound electron-hole pair we have obtained a system of discrete hydrogen-like energy levels. We, however, failed to take account of the motion of the exciton as a whole, i. e. of its inertia centre. Kinetic energy of this motion is

$$T^{ex} = \frac{\hbar^2 |\mathbf{K}^{ex}|^2}{2(m_n^* + m_p^*)}, \tag{27.11}$$

and it should be added to E_n^{ex}. Thus, each energy level E_n^{ex} should grow into a rather wide band—*the exciton band*. However, only comparatively narrow light absorption lines attributed to excitons are observed experimentally. The reason for this should be sought in the selection rules for the wave vector.

In conclusion we would like to mention an excited state of another type which is typical of substances with a large amount of ionic bonding. This is the polaron the essence of which may be pictured as follows. A conduction electron establishes an electron field which polarizes the surrounding space. This resultant induced positive charge interacts with the electron to create additional potential energy in the form of a potential trough which localizes the electron. The polarized space travels with the electron. *The system consisting of an electron and a volume of space polarized by its field is termed polaron.*

Summary of Secs. 26-27

1. The calculation of the band pattern of specific semiconductors is a complicated problem. The dependence $E(\mathbf{k})$ may be found only for some parts of the Brillouin zone. Calculations have been

carried out for germanium and silicon. They serve as a basis for calculating the band pattern of $A^{III}B^{V}$ and $A^{II}B^{VI}$ compounds. Main source of information about the band pattern is the analysis of experimental data.

2. Energy extrema for electrons and holes in germanium and silicon are at different points of the Brillouin zone. In $A^{III}B^{V}$ compounds they might be both at different points and at one point.

3. The valence bands of germanium and silicon are made up of three branches of E(k) curve each of which has its own type of holes. The constant-energy surfaces are of the shape of deformed spheres (corrugated surfaces).

4. The conduction band of germanium has eight equivalent minima lying on the Brillouin zone boundary in the direction [111]. Constant-energy surfaces are rotational ellipsoids the rotational axes of which coincide with the [111] direction. There are four full ellipsoids to the first Brillouin zone.

The conduction band of silicon has six equivalent minima lying near the Brillouin zone boundary in the [100] direction. The constant-energy surfaces are rotational ellipsoids whose rotational axes lie in the [100] direction.

5. All conclusions of the band theory relating to the dependence of the forbidden band width on the composition of the substance, to the relation between the electron effective mass and the type of the band, etc. are in good qualitative agreement with experiment.

6. The band pattern of the energy spectrum follows both from long- and short-distance order.

7. All results of the band theory are actually applicable to quasiparticles which serve as the bearers of specific properties reflecting the properties of the solid.

Chapter III

ELECTRON AND HOLE STATISTICS IN SEMICONDUCTORS

28. DENSITY OF STATES

The conductivity of a substance is determined by the concentration and mobility of charge carriers, therefore, in order to understand the effect of the ambient on conductivity, we will have to elucidate how the concentration and mobility depend on the ambient, first of all on temperature.

We will consider free charge carriers, i. e. electrons in a conduction band and holes in a valence band. To calculate the number of charge carriers one must' know the number of states and the probability of charge carriers occupying these states.

Suppose the probability of electrons occupying a unit volume of the phase space with the centre at a point (**r**, **k**) at the moment t is f (**r**, **k**, t); we regard **r** and **k** (or r and **P**) as canonically conjugate quantities satisfying the usual Heisenberg commutation relations

$$r_l P_m - P_m r_l = i\hbar\delta_{lm}; \qquad (l, m = x, y, z). \tag{28.1}$$

There are $\frac{d\Gamma}{h^3}$ *phase cells in the phase space volume element*, $d\Gamma$ *containing* $2\,\frac{d\Gamma}{h^3}$ *states* (the coefficient 2 shows that there can be two electrons with opposite spins in each cell). *The number of electrons* dn *in the volume* $d\Gamma$ *is equal to the product of the number of states by the probability of electrons occupying these states:*

$$dn = dn\,(\mathbf{r}, \mathbf{k}, t) = f\,(\mathbf{r}, \mathbf{k}, t)\,2\,\frac{d\Gamma}{h^3}. \tag{28.2}$$

$d\Gamma$ can be expressed in terms of the volumes of the geometrical $d\tau_r$ and quasimomentum $d\tau_p = \hbar^3\,d\tau_k$ spaces:

$$d\Gamma = d\tau_r d\tau_p = dx\,dy\,dz\,dk_x\,dk_y\,dk_z\hbar^3. \tag{28.3}$$

Integrating over the volumes of the crystal V and of the first Brillouin zone V_k we obtain the full number of electrons N in an energy band of the crystal:

$$N=\frac{2}{h^3}\iint\limits_{(VV_{\mathbf{k}})} f(\mathbf{r},\mathbf{k},t)\,d\Gamma. \qquad (28.4)$$

Mean electron concentration is:

$$\bar{n}=\frac{N}{V}=\frac{2}{Vh^3}\iint\limits_{(VV_{\mathbf{k}})} f(\mathbf{r},\mathbf{k},t)\,d\Gamma=\frac{1}{4\pi^3V}\iint\limits_{(VV_{\mathbf{k}})} f(\mathbf{r},\mathbf{k},t)\,d\tau_{\mathbf{r}}\,d\tau_{\mathbf{k}}. \qquad (28.5)$$

The number of electrons $dn_{\mathbf{r}}$ in an element of volume $d\tau_{\mathbf{r}}$ is:

$$dn_{\mathbf{r}}=n(\mathbf{r},t)\,d\tau_{\mathbf{r}}=d\tau_{\mathbf{r}}\cdot\frac{2}{h_3}\int\limits_{(V_{\mathbf{k}})} f(\mathbf{r},\mathbf{k},t)\,d\tau_{\mathbf{p}}. \qquad (28.6)$$

In other words, electron concentration is

$$n,(\mathbf{r},t)=\frac{dn_{\mathbf{r}}}{d\tau_{\mathbf{r}}}=\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})} f(\mathbf{r},\mathbf{k},t)\,d\tau_{\mathbf{k}}. \qquad (28.7)$$

$V_{\mathbf{k}}$ defines the volume of the first Brillouin zone no matter whether it is in the **k**- or **P**-space. For a uniform crystal the probability f is independent of co-ordinates, and integration over $d\tau_{\mathbf{r}}$ yields, therefore, the volume of the crystal. As a result $\bar{n}$ of (28.5) is equal to the actual concentration n of (28.7):

$$\bar{n}=\frac{1}{4\pi^3V}\iint\limits_{(VV_{\mathbf{k}})} f(\mathbf{k},t)\,d\tau_{\mathbf{r}}\,d\tau_{\mathbf{k}}=\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})} f(\mathbf{k},t)\,d\tau_{\mathbf{k}}=n(t). \qquad (28.8)$$

As was stated in the first chapter the distribution function is a function of energy and temperature. For stationary states it is independent of time. Since energy is an eigenvalue of the Hamilton operator for a quantum system it is independent of the co-ordinates, by force of which the distribution function $f_0=f_0(E;T)$, where $f_0(E,T)$ is the Fermi-Dirac (or Maxwell-Boltzmann) function, will be independent of co-ordinates, too.

In the energy interval dE there may be various numbers of states dS (spin not taken into account) per unit volume of the crystal depending on the energy.

Suppose an equation connects dS and dE:

$$dS=dS(E)=\mathcal{N}(E)\,dE. \qquad (28.9)$$

$N(E)$ is the number of states per unit energy interval per unit crystal volume. $N(E)$ is termed density of states:

$$N(E)=\frac{dS}{dE}. \qquad (28.10)$$

If $f_0(E, T)$ is the probability of electrons occupying states with the energy E then the number of electrons in these states dn will be equal to

$$dn = dn\,(E, T) = 2f_0\,(E, T)\,dS = 2f_0\,(E, T)\,N\,(E)\,dE. \quad (28.11)$$

Electron concentration will be

$$n = n\,(T) = 2\int\limits_{(E)} f_0\,(E, T)\,N\,(E)\,dE. \quad (28.12)$$

The integration should be performed from the bottom to the top of the conductivity band E_c. Since $f_0(E, T)$ sharply depends on energy, the upper limit may be put at infinity, and correspondingly

$$n = 2\int\limits_{E_c}^{\infty} f_0\,(E, T)\,N\,(E)\,dE. \quad (28.13)$$

The quantity $N(E)$ is closely related to the shape of constant-energy surfaces.

Indeed, build two constant-energy surfaces E and $E+dE$ in the Brillouin zone. They cut out a layer in the quasimomentum space (Fig. 45). Let the volume of this layer be $d\tau_p$ and the corresponding volume of the phase space

$$d\Gamma = V\,d\tau_p\,. \quad (28.14)$$

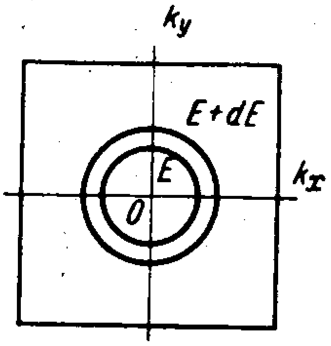


Fig. 45. The volume of a layer inside the Brillouin zone bounded by the constant-energy surfaces E and $E+dE$

The number of phase cells in this volume is $\frac{d\Gamma}{h^3}$, and the number of states per unit volume (geometrical) $dS = \frac{d\Gamma}{Vh^3}$. On the other hand, $dS = N\,(E)\,dE$, therefore

$$dS = N\,(E)\,dE = \frac{1}{V}\,\frac{d\Gamma}{h^3} = \frac{d\tau_p}{h^3} = \frac{d\tau_k}{4\pi^3} \quad (28.15)$$

The quantity $d\tau_k\,(E, E+dE)$ of (28.15) may be found if the equation for constant-energy surfaces is known. Note that the quantities

$$\frac{1}{h^3},\quad \frac{1}{Vh^3}\ \text{and}\ \frac{1}{V\cdot 8\pi^3} \quad (28.16)$$

may be regarded as densities of states in the phase space, in the quasimomentum space and in the wave vector space, respectively.

Consider some specific cases.

1. **Spherical constant-energy surfaces with E_{min} in the centre of the Brillouin zone.** Let

$$E\,(\mathbf{k}) = E_c + \frac{\hbar^2 k^2}{2m^*}\,. \quad (28.17)$$

Two constant-energy surfaces E and $E+dE$ cut out a spherical layer dk thick with a volume $d\tau_{\mathbf{k}}(E, E+dE)$ (Fig.45):

$$d\tau_{\mathbf{k}} = 4\pi k^2\, dk. \qquad (28.18)$$

Express k in terms of E

$$k^2 = \frac{2m^*}{\hbar^2}(E-E_c); \quad k = \left(\frac{2m^*}{\hbar^2}\right)^{1/2}(E-E_c)^{1/2}. \qquad (28.19)$$

Differentiating the first relation of (28.19) with respect to k we obtain

$$2k\, dk = \frac{2m^*}{\hbar^2}dE. \qquad (28.20)$$

Taking into account (28.18), (28.19) and (28.20) we may write

$$d\tau_{\mathbf{k}}(E,\ E+dE) = 4\pi k^2\, dk = 2\pi k\cdot 2k\, dk =$$
$$= 2\pi\left(\frac{2m^*}{\hbar^2}\right)^{1/2}(E-E_c)^{1/2}\cdot\frac{2m^*}{\hbar^2}dE = 2\pi\left(\frac{2m^*}{\hbar^2}\right)^{3/2}(E-E_c)^{1/2}dE. \qquad (28.21)$$

From (28.15) and (28.21) we obtain for dS

$$dS = \frac{d\tau_{\mathbf{k}}}{8\pi^3} = N(E)\, dE = \frac{2\pi\,(2m^*)^{3/2}}{8\pi^3\hbar^3}(E-E_c)^{1/2}dE, \qquad (28.22)$$

or

$$N(E) = 2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}(E-E_c)^{1/2}. \qquad (28.23)$$

2. **Spherical constant-energy surfaces with the energy minimum at points $\mathbf{k}_0$.** Let the energy minima, M in number, be not in the centre of Brillouin zone, but at some points $\mathbf{k}_0$. Build constant-energy surfaces E and $E+dE$ to obtain M spheres (Fig. 46). The equation for one of them is of the form

$$E = E_c + \frac{\hbar^2(\mathbf{k}-\mathbf{k}_0)^2}{2m^*}. \qquad (28.24)$$

The radii of the spheres are $|\mathbf{k}-\mathbf{k}_0|$, the thickness of the spherical layer $dk = d|\mathbf{k}-\mathbf{k}_0|$; but the number of the spheres is now M, therefore M layers correspond to the energy interval dE

$$d\tau_{\mathbf{k}}(E,\ E+dE) = M\cdot 4\pi(\mathbf{k}-\mathbf{k}_0)^2 d|\mathbf{k}-\mathbf{k}_0|. \qquad (28.25)$$

Expressing $|\mathbf{k}-\mathbf{k}_0|$ and $d|\mathbf{k}-\mathbf{k}_0|$ in terms of energy we obtain

$$N(E) = M\cdot 2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}(E-E_c)^{1/2}. \qquad (28.26)$$

3. **Ellipsoidal constant-energy surfaces.** Consider a more general case when the constant-energy surfaces have the shape of an ellipsoid:

$$E(\mathbf{k}) = E_c + \frac{\hbar^2}{2}\left(\frac{k_x^2}{m_1} + \frac{k_y^2}{m_2} + \frac{k_z^2}{m_3}\right), \tag{28.27}$$

or, in a canonical form,

$$\frac{k_x^2}{a^2} + \frac{k_y^2}{b^2} + \frac{k_z^2}{c^2} = 1, \tag{28.28}$$

where the semi-axes of the ellipsoid are

$$a_i = \left[\frac{2m_i(E-E_c)}{\hbar^2}\right]^{1/2}. \tag{28.29}$$

The volume of one ellipsoid with the semi-axes a, b, c is equal to

$$\tau_{\mathbf{k}}^{(1)} = \frac{4\pi}{3}abc = \frac{4\pi}{3\hbar^3}(8m_1m_2m_3)^{1/2}(E-E_c)^{3/2}, \tag{28.30}$$

and the volume of one layer between two ellipsoidal constant-energy surfaces

$$d\tau_{\mathbf{k}}^{(1)} = \frac{2\pi}{\hbar^3}(8m_1m_2m_3)^{1/2}(E-E_c)^{1/2}dE. \tag{28.31}$$

Substituting $d\tau_{\mathbf{k}}^{(1)}$ into the expression for dS and taking account of the fact that there may be M energy minima, we obtain

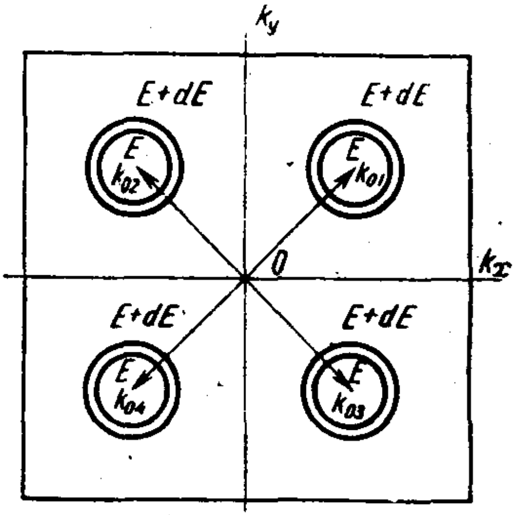


Fig. 46. The volume of a layer inside the Brillouin zone corresponding to the energy interval dE for the case of M extrema

$$dS = \frac{M d\tau_{\mathbf{k}}^{(1)}}{8\pi^3} = \frac{M\cdot 2\pi}{h^3}(8m_1m_2m_3)^{1/2}(E-E_c)^{1/2}dE, \tag{28.32}$$

or

$$N(E) = \frac{M\cdot 2\pi}{h^3}(8m_1m_2m_3)^{1/2}(E-E_c)^{1/2}. \tag{28.33}$$

The reader should be reminded that, for instance, for germanium $M=4$, and for silicon $M=6$.

Thus, *the density of states is proportional to* $(E-E_c)^{1/2}$ and to $(m_1 m_2 m_3)^{1/2}$, where m_1, m_2, m_3 are the components of the effective mass tensor (in a diagonal tensor). $N(E)$ *will be small for small* m_i. Obviously, $N(E)$ should be small for intermetallic compounds.

If one puts

$$M^2 (m_1 m_2 m_3) = m_d^{*3}, \tag{28.34}$$

where m_d^* is termed effective mass for the density of states, then generally

$$N(E) = 2\pi \left(\frac{2m_d^*}{h^2}\right)^{3/2} (E-E_c)^{1/2}. \tag{28.35}$$

This looks exactly like (28.23) for the case of an isotropic mass with one minimum.

The expression (28.35) is the most general since (28.23), (28.26) follow from it as specific cases. It should, however, be remembered that *the dependence of* $N(E)$ *on energy of the form* $N(E) \sim \sim (E-E_c)^{1/2}$ *is valid only insofar as energy remains a quadratic function of quasimomentum*, in other words, *the expression* (28.35) *is valid only for states near the energy minimum, i.e. at the bottom of the band.*

Find the expression for the density of states near the top of energy band where, too, the energy is a quadratic function of quasimomentum. If energy maxima, M in number, are at points $\mathbf{k}_0$ of the Brillouin zone then $E(\mathbf{k})$ (for each maximum) may be written in a general form

$$E(\mathbf{k}) = E(\mathbf{k}_0) + \frac{\hbar^2}{2}\left[\frac{(k_x-k_{0x})^2}{m_1} + \frac{(k_y-k_{0y})^2}{m_2} + \frac{(k_z-k_{0z})^2}{m_3}\right]. \tag{28.36}$$

Since the electron effective mass tensor is negative in the energy maximum, we may introduce instead a hole effective mass tensor satisfying the condition $m_{ip} = -m_{in}$ and write

$$E(\mathbf{k}) = E(\mathbf{k}_0) - \frac{\hbar^2}{2}\left[\frac{(k_x-k_{0x})^2}{m_{1p}} + \frac{(k_y-k_{0y})^2}{m_{2p}} + \frac{(k_z-k_{0z})^2}{m_{3p}}\right], \tag{28.37}$$

or

$$\frac{\hbar^2}{2}\left[\frac{(k_x-k_{0x})^2}{m_{1p}} + \frac{(k_y-k_{0y})^2}{m_{2p}} + \frac{(k_z-k_{0z})^2}{m_{3p}}\right] = E_v - E(\mathbf{k}), \tag{28.38}$$

where $E_v = E(\mathbf{k}_0)$ is, in this case, the energy of the top of the valence band. Omitting obvious calculations we obtain *the expression for the density of states near the top of the band*

$$N(E) = 2\pi \left(\frac{2m_{pd}^*}{h^2}\right)^{3/2} (E_v - E)^{1/2}, \tag{28.39}$$

where m^*_{pd} *denotes the hole effective mass for the density of states:*

$$m^*_{pd} = (M^2 m_{p1} m_{p2} m_{p3})^{1/3}. \qquad (28.40)$$

The densities of states near the bottom and the top of the band are exactly of the same form since the expressions obtained for $N(E)$ follow from the quadratic dependence of energy on quasimomentum in the vicinity of the extrema. The difference in signs

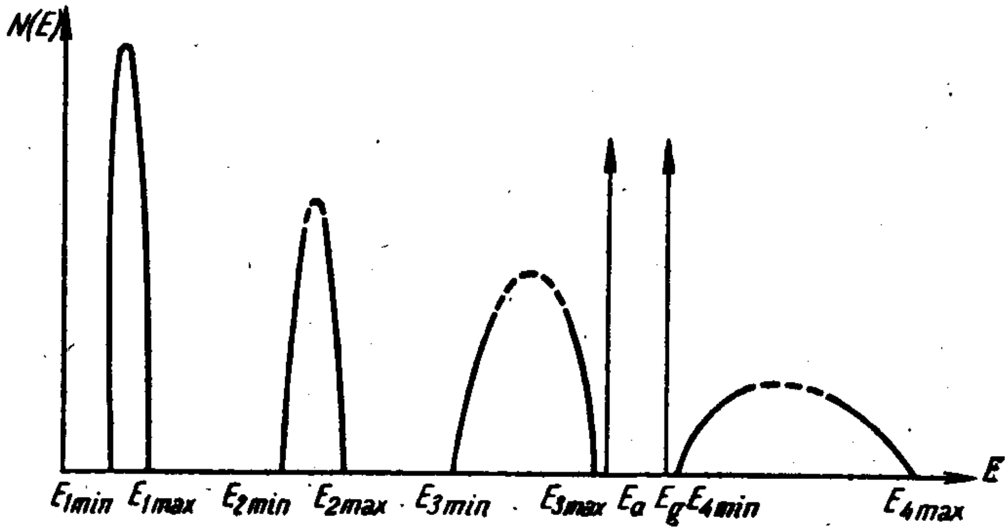


Fig. 47. The density of states in different bands and on the localized energy levels .

of the expressions under the radical sign ($E-E_c$ and E_v-E) is due to the difference in signs of effective masses in the minimum and maximum of energy. One of the expressions for $N(E)$ remains true as long as quadratic dependence of energy on quasimomentum holds. Away from boundaries the expression for $N(E)$ cannot be written unless the dependence $E(\mathbf{k})$ is known. The total number of states in a band with account taken of spins is $2N$ or $2Ng$ (g is the degeneracy factor), therefore

$$2\int_{E_{min}}^{E_{max}} N(E)\,dE = 2Ng. \qquad (28.41)$$

Figure 47 shows the general shape of the function $N(E)$. For forbidden bands $N(E)=0$ (if localized electron states are ignored).

To find the density of localized states one can make use of the following simple arguments. Each impurity atom can have one electron and so the total number of states will be N_d and N_a, respectively.

If one assumes that there is no impurity band, and impurity levels remain discrete the density of states for each level must be taken as infinite; however, the integral of the density over all possible states must be unity (for a single level).

To this end the density of states should be expressed with the aid of the δ-function:

$$N_d(E) = N_d\delta(E - E_d), \tag{28.42}$$

$$N_a(E) = N_a\delta(E - E_a), \tag{28.43}$$

where $N_d(E)$, $N_a(E)$ denote the density of states at the donor and the acceptor levels having the energy E_d and E_a, respectively. The density of states $N_d(E)$ and $N_a(E)$ is shown in Fig. 47 by a vertical line. When the increase in impurity concentration leads to the formation of an impurity band the expressions (28.42) and (28.43) cease to be valid and must be superseded by the expressions (28.35) or (28.39).

Consider a general method of calculating the density of states $N(E)$. The essence of it is that the integration over the Brillouin zone volume is done in two steps—at first one integrates over the constant-energy surface, and then over the energy. This method is convenient in the case of energy dependent quantities under the integral sign assuming constant values on the constant-energy surface. The volume element $d\tau_{\mathbf{k}}$ may be expressed in terms of the surface element dS_E and the normal component of the wave vector dk_n

$$d\tau_{\mathbf{k}} = dS_E\, dk_n \tag{28.44}$$

which, in its turn, is expressed in terms of dE:

$$dE = \left(\frac{dE}{d\mathbf{k}} \cdot d\mathbf{k}\right) = |\nabla_{\mathbf{k}} E|\, dk_n \tag{28.45}$$

and

$$dk_n = \frac{dE}{|\nabla_{\mathbf{k}} E|}; \quad d\tau_{\mathbf{k}} = \frac{dS_E \cdot dE}{|\nabla_{\mathbf{k}} E|}. \tag{28.46}$$

As a result we obtain for n

$$n = \frac{1}{4\pi^3} \int f \frac{dS_E\, dE}{|\nabla_{\mathbf{k}} E|}. \tag{28.47}$$

Since the distribution function depends on energy, it follows:

$$n = 2 \int_{(E)} f(E, T)\, dE \int_{(S_E)} \frac{dS_E}{8\pi^3 |\nabla_{\mathbf{k}} E|}. \tag{28.48}$$

Comparing (28.48) with (28.12) we obtain for the density of states:

$$N(E) = \frac{1}{8\pi^3} \int_{(S_E)} \frac{dS_E}{|\nabla_{\mathbf{k}} E|}. \tag{28.49}$$

From (28.49) it follows for spherical constant-energy surfaces:

$$N(E) = \frac{1}{8\pi^3 |\nabla_{\mathbf{k}} E|} \cdot 4\pi \mathbf{k}^2 = \frac{\mathbf{k}^2}{2\pi^2 |\nabla_{\mathbf{k}} E|}. \tag{28.50}$$

For a quadratic dispersion law, (28.50) leads to (28.23) and all that follows from it. However, for a non-quadratic dispersion law the relation (28.49) is a more appropriate one.

Let us find the expression for the density of states using this relation and the Kane dispersion law (26.37). Taking the bottom of the conductivity band as the origin of the energy scale we obtain after subtracting E_g from (26.37)

$$E(\mathbf{k}) = -\frac{E_g}{2} + \frac{\hbar^2\mathbf{k}^2}{2m} + \sqrt{\frac{2\mathbf{P}^2\mathbf{k}^2}{3} + \frac{E_g^2}{4}}. \qquad (28.51)$$

Since $\mathbf{P} \approx 6\cdot 10^{-8}$ eV·cm the second term may be neglected for small **k**, and we may write

$$\left[E(\mathbf{k}) + \frac{E_g}{2}\right]^2 = \frac{2\mathbf{P}^2\mathbf{k}^2}{3} + \frac{E_g^2}{4} \qquad (28.52)$$

or

$$\frac{2\mathbf{P}^2\mathbf{k}^2}{3} = E(\mathbf{k})\left[E_g + E(\mathbf{k})\right]. \qquad (28.53)$$

In general, it follows from (28.51) that

$$|\nabla_{\mathbf{k}} E| = \frac{dE}{dk} = \frac{\hbar^2\mathbf{k}}{m} + \frac{(4\mathbf{P}^2/3)\,k}{\sqrt{\frac{2}{3}\,\mathbf{P}^2\mathbf{k}^2 + \frac{E_g^2}{4}}} \qquad (28.54)$$

or, when the term $\frac{\hbar^2 k}{m}$ is neglected

$$|\nabla_{\mathbf{k}} E| = \frac{\frac{4\mathbf{P}^2 k}{3}}{\left(\frac{2\mathbf{P}^2\mathbf{k}^2}{3} + \frac{E_g^2}{4}\right)^{1/2}}. \qquad (28.55)$$

Substituting (28.55) into (28.50) we obtain

$$N(E) = \frac{3k\sqrt{\frac{2}{3}\,\mathbf{P}^2\mathbf{k}^2 + \frac{E_g^2}{4}}}{2\pi^2\cdot 4\mathbf{P}^2}. \qquad (28.56)$$

Cancelling out **k** we obtain, in compliance with (28.53), (28.56) in the form

$$\mathrm{V}(E) = \frac{E^{1/2}\,(E_g + E)^{1/2}\left(E + \frac{E_g}{2}\right)}{4\pi^2\left(\frac{\mathbf{P}^2\cdot 2}{3}\right)^{3/2}}. \qquad (28.57)$$

Introduce the effective mass of the bottom of the conductivity band m_n^{0*} by the relation

$$m_n^{0*} = \frac{3\hbar^2 E_g}{4\mathbf{P}^2}. \qquad (28.58)$$

This enables us to write (28.57), by analogy with expressions previously used, in the form

$$N(E)=\frac{1}{2}\left(\frac{2m_n^{0*}}{\hbar^2}\right)^{3/2}E^{1/2}\left(1+\frac{E}{E_g}\right)^{1/2}\left(1+2\frac{E}{E_g}\right). \tag{28.59}$$

When $E_g=0$ it is easier to derive the expression for $N(E)$ from (28.57)

$$N(E)=\frac{E^2}{\left(\frac{4\pi \mathrm{P}^2}{3}\right)^{3/2}}. \tag{28.60}$$

For $\mathrm{P}\approx 6\times 10^{-8}$ eV·cm we have $N(E)(\mathrm{cm}^{-3})\approx 5.4\times 10^{30}[E(\mathrm{eV})]^2$.

29. ELECTRON AND HOLE CONCENTRATIONS

Now let us write the expression (28.13) for electron concentration substituting the expression (28.35) for $N(E)$. We may do this because the Fermi-Dirac function falls off very rapidly with the increase in energy, and the contribution of the electrons occupying upper energy states to the electron concentration is, accordingly, negligible; so that the use of (28.35) for $N(E)$ for all energy values does not lead to noticeable errors in n. To illustrate this point a graph of the product of $f_0(E, T)$ by $N(E)$ in the form of (28.35) for a metal ($E_c=0$; $F>0$) is shown in Fig. 48.

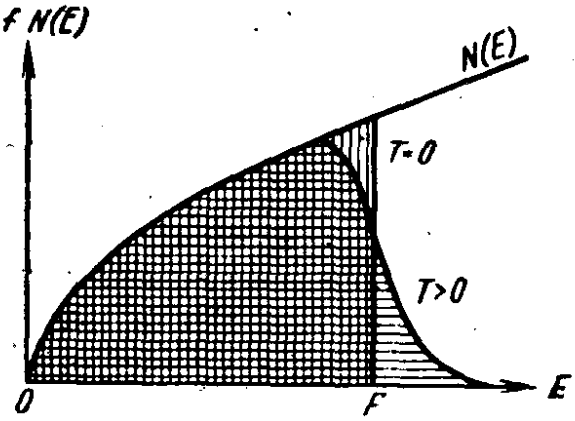


Fig. 48. The graph of the product of $N(E)$ and $f(E, T)$ for metals and degenerate semiconductors

Substituting the expression (28.35) for $N(E)$ and the equilibrium state Fermi-Dirac function

$$f_0(E, T)=\frac{1}{e^{\frac{E-F}{kT}}+1} \tag{29.1}$$

into (28.13) we obtain

$$n=2\int_{E_c}^{\infty} f_0(E, T)\,N(E)\,dE=2\int_{E_c}^{\infty}\frac{2\pi\left(\frac{2m_d^*}{h^2}\right)^{3/2}(E-E_c)^{1/2}dE}{e^{\frac{E-F}{kT}}+1}. \tag{29.2}$$

The expression (29.2) enables the electron concentration to be calculated provided F is known.

Adopting dimensionless variables we re-write the expression (29.2) for n setting:

$$\frac{E-E_c}{kT}=x; \quad dx=\frac{dE}{kT}; \quad \frac{F-E_c}{kT}=\xi. \tag{29.3}$$

Hence

$$n=4\pi\left(\frac{2m_d^*kT}{h^2}\right)^{3/2}\int_0^\infty \frac{x^{1/2}\,dx}{e^{x-\xi}+1}. \tag{29.4}$$

Introduce the following notation:

$$2\left(\frac{2\pi m_d^*kT}{h^2}\right)^{3/2}=N_c, \tag{29.5}$$

$$\int_0^\infty \frac{x^{1/2}\,dx}{e^{x-\xi}+1}=\Phi_{1/2}(\xi). \tag{29.6}$$

The quantity N_c is termed effective number of states in the conduction band, and $\Phi_{1/2}(\xi)$—Fermi integral of the order of $\frac{1}{2}$. In new notation the electron concentration n assumes the form

$$n=\frac{2N_c}{\sqrt{\pi}}\Phi_{1/2}(\xi). \tag{29.7}$$

Electron concentration is a function of temperature and the Fermi level $n=n(T, F)$.

Generally, the Fermi integral $\Phi_{1/2}(\xi)$ cannot be expressed in terms of elementary functions; however, for a number of practically important applications there are approximate analytical expressions which will be derived below.

Now let us find the expression for the number of free holes in the valence band. The density of states near the top of the valence band is given by the expression (28.39). The hole distribution function is:

$$f_{0p}(E, T)=1-f_0(E, T)=\frac{1}{e^{\frac{F-E}{kT}}+1}. \tag{29.8}$$

Accordingly, the number of holes dp in the energy interval dE will be equal to the number of states $2N(E)\,dE$ multiplied by the probability of their being occupied by holes f_{0p}:

$$dp=2N(E)\,f_{0p}(E, T). \tag{29.9}$$

The hole concentration in the valence band is

$$p=2\int_{-\infty}^{E_v} f_{0p}(E, T)\,N(E)\,dE. \tag{29.10}$$

We substituted $-\infty$ for the lower limit $E_{v\,min}$ because of the steep dependence of f_0 on energy. Substituting the expressions (29.8) for f_{0p} and (28.39) for $N(E)$ we obtain

$$p=2\int_{-\infty}^{E_v} 2\pi\left(\frac{2m_{pd}^*}{h^2}\right)^{3/2}\frac{(E_v-E)^{1/2}\,dE}{e^{\frac{F-E}{kT}}+1}. \qquad (29.11)$$

The energy scale for holes in (29.9-11) and subsequent expressions coincides with that for electrons.

Introduce the notation:

$$\frac{E_v-E}{kT}=x;\quad dx=-\frac{dE}{kT};\quad \frac{E_v-F}{kT}=\eta, \qquad (29.12)$$

and re-write (29.11) as

$$p=4\pi\left(\frac{2m_{pd}^*kT}{h^2}\right)^{3/2}\int_0^{\infty}\frac{x^{1/2}\,dx}{e^{x-\eta}+1}=\frac{2N_v}{\sqrt{\pi}}\,\Phi_{1/2}(\eta), \qquad (29.13)$$

where

$$N_v=2\left(\frac{2\pi m_{pd}^*kT}{h^2}\right)^{3/2} \qquad (29.14)$$

is *the effective number of states in the valence band*, and

$$\Phi_{1/2}(\eta)=\int_0^{\infty}\frac{x^{1/2}\,dx}{e^{x-\eta}+1} \qquad (29.15)$$

is the Fermi integral of the form identical with that of $\Phi_{1/2}(\xi)$.

The electron n and hole p concentrations in compliance with (29.7) *and* (29.13) *depend on temperature T and Fermi level F.*

There are various approximate expressions for the Fermi integral $\Phi_{1/2}$, each one valid in a corresponding range of the argument, i.e.

$$\Phi_{1/2}(\xi)=\begin{cases}\dfrac{\sqrt{\pi}}{2}e^{\xi} \text{ for } -\infty<\xi<-1\\[2mm] \dfrac{\sqrt{\pi}}{2}\dfrac{1}{0.25+e^{-\xi}} \text{ for } -1<\xi<5,\\[2mm] \dfrac{2}{3}\xi^{3/2} \text{ for } 5<\xi<\infty.\end{cases} \qquad (29.16)$$

The first approximation valid for $\xi<-1$ *corresponds to Boltzmann statistics. The condition for classical statistics to be valid is the inequality* $\xi<-1$, or $\frac{F-E_c}{kT}<-1$, *whence* $F<E_c-kT$,

i.e. the semiconductor is non-degenerate (obeys the laws of classical statistics) *when the Fermi level lies below the bottom of the conduction band at a distance greater than kT. If the Fermi level is more than 5 kT above E_c the semiconductor will be completely degenerate.* For the intermediate case $E_c - kT < F < E_c + 5kT$ the properties of the semiconductor are transitional from those of a non-degenerate to those of a completely degenerate one. As may be seen from the above, the condition of degeneracy includes the temperature and the position of the Fermi level with respect to the bottom of the conduction band.

Relations will be derived below linking this condition with other semiconductor parameters, first of all with impurity concentration.

Let us show that the above approximate expressions for $\Phi_{1/2}(\xi)$ actually hold.

It follows directly from the expression for $\Phi_{1/2}(\xi)$ that for $\xi < -1$ the exponent will exceed unity ($e^{-\xi} > 1$) for all $x > 0$. This means that the Fermi-Dirac function may be replaced by the Boltzmann function for all $\xi < -1$:

$$\frac{1}{e^{x-\xi}+1} \cong e^{-x+\xi} \text{ (for } \xi < -1;\ x > 0). \tag{29.17}$$

Hence

$$\Phi_{1/2}(\xi) \cong \int_0^\infty x^{1/2} e^{-x+\xi}\, dx = e^{\xi} \int_0^\infty x^{1/2} e^{-x}\, dx. \tag{29.18}$$

The last integral in (29.18) is the Euler gamma-function $\Gamma(^3/_2)$ (which may be reduced to the Poisson integral):

$$\Gamma(^3/_2) = \int_0^\infty x^{1/2} e^{-x}\, dx = \frac{\sqrt{\pi}}{2}, \tag{29.19}$$

and

$$\Phi_{1/2}(\xi) \cong \frac{\sqrt{\pi}}{2} e^{\xi} \tag{29.20}$$

in full agreement with (29.16).

Write the expression for electron concentration in a non-degenerate semiconductor

$$n = \frac{2N_c}{\sqrt{\pi}} \Phi_{1/2}(\xi) \cong N_c e^{\xi} = N_c e^{-\frac{E_c - F}{kT}} =$$

$$= 2\left(\frac{2\pi m_d^* kT}{h^2}\right)^{3/2} e^{-\frac{E_c - F}{kT}}; \quad F < E_c - kT. \tag{29.21}$$

The effective number of states is temperature dependent. Substituting numerical values of the universal constants which enter the expression for N_c we obtain

$$N_c = 2\left(\frac{2\pi m_d^* kT}{h^2}\right)^{3/2} = 4.82\cdot 10^{15}\left(\frac{m_d^*}{m}\right)^{3/2} T^{3/2} =$$

$$= 2.5\cdot 10^{19}\left(\frac{m_d^*}{m}\right)^{3/2}\left(\frac{T}{300}\right)^{3/2}, \tag{29.22}$$

$$n = 4.82\cdot 10^{15}\left(\frac{m_d^*}{m}\right)^{3/2} T^{3/2} e^{-\frac{E_c-F}{kT}} \tag{29.23}$$

Consider now the case of a completely degenerate semiconductor for which

$$\Phi_{1/2}(\xi) = \frac{2}{3}\xi^{3/2} = \frac{2}{3}\left(\frac{F-E_c}{kT}\right)^{3/2} \quad (F > E_c + 5kT). \tag{29.24}$$

The electron concentration will be

$$n = \frac{2N_c}{\sqrt{\pi}}\Phi_{1/2}(\xi) = \frac{8\pi}{3}\left(\frac{2m_d^*}{h^2}\right)^{3/2}(F-E_c)^{3/2}, \tag{29.25}$$

i.e. it will be independent of temperature. One should keep in mind that the Fermi level in this case is in the conduction band more than $5kT$ above its bottom.

Find the conditions when the Fermi integral may be represented in the form $\Phi_{1/2}(\xi) = \frac{2}{3}\xi^{3/2}$. For low temperatures, as is well known, $\left(-\frac{\partial f_0}{\partial E}\right) \cong \delta(E-F)$. This approximation enables the integrals of the product of two functions $\varphi(E) f_0(E, T)$ to be easily calculated.

Indeed,

$$\int_0 \varphi(E) f_0(E, T)\, dE = X(E) f_0(E, T)\Big|_0^\infty - \int_0^\infty X(E)\frac{\partial f_0}{\partial E} dE =$$

$$= -X(0) f_0(0, T) + X(F) \approx X(F) - X(0), \tag{29.26}$$

where

$$X(E) = \int \varphi(E)\, dE, \text{ and } f_0(0, T) \approx 1. \tag{29.27}$$

Making use of this property of Fermi-Dirac distribution we may easily obtain the expression for the concentration. In this case

$\varphi(E)=(E-E_c)^{1/2}$ and $X(E)=\frac{2}{3}(E-E_c)^{3/2}$, therefore

$$n=\frac{2N_c}{\sqrt{\pi}}\Phi_{1/2}(\xi)=\frac{2N_c}{\sqrt{\pi}\,(kT)^{3/2}}\int\limits_{E_c}^{\infty}\frac{(E-E_c)^{1/2}\,dE}{e^{\frac{E-F}{kT}}+1}=$$

$$=\frac{2N_c}{\sqrt{\pi}\,(kT)^{3/2}}\cdot\frac{2}{3}(F-E_c)^{3/2}. \qquad (29.28)$$

It follows from the expression (29.28) that, indeed

$$\Phi_{1/2}(\xi)=\frac{2}{3}\left(\frac{F-E_c}{kT}\right)^{3/2}=\frac{2}{3}\xi^{3/2}. \qquad (29.29)$$

The condition for the independence of n from temperature is tantamount to the condition that $-\frac{\partial f_0}{\partial E}\cong\delta(E-F)$. This, ~~however,~~ entails the maximum possible variation rate of $f_0(E,$ ~~$T)$ in~~ the vicinity of $E\cong F$, because the narrower is the area $(E-F)$ where $f_0(E, T)$ changes, the closer does its derivative resemble the δ-function, the better does the δ-nature of $-f_0'(E, T)$ reveal itself in the low temperature range.

In the transition interval from the non-degenerate to completely degenerate case the temperature dependence of the electron concentration n will be:

$$n=N_c\frac{1}{0.25+e^{\frac{E_c-F}{kT}}}. \qquad (29.30)$$

The results obtained for electrons may be easily applied to holes. It suffices to this end to re-write the expression for $\Phi_{1/2}(\eta)$ in the form of (29.16):

$$\Phi_{1/2}(\eta)=\begin{cases}\frac{\sqrt{\pi}}{2}e^{\eta} \text{ for } -\infty<\eta<-1,\\ \frac{\sqrt{\pi}}{2}\frac{1}{0.25+e^{-\eta}} \text{ for } -1<\eta<5,\\ \frac{2}{3}\eta^{3/2} \text{ for } 5<\eta<\infty.\end{cases} \qquad (29.31)$$

In a non-degenerate semiconductor the hole concentration is governed by the Boltzmann statistics the condition of applicability of which is the inequality

$$\eta=\frac{E_v-F}{kT}<-1, \text{ or } F>E_v+kT, \qquad (29.32)$$

i.e. *the Fermi level in a non-degenerate semiconductor must lie at least by the amount kT above the top of the valence band.*

For a completely degenerate semiconductor:

$$\eta = \frac{E_v - F}{kT} > 5, \text{ or } F < E_v - 5kT. \qquad (29.33)$$

i.e. *the Fermi level in a completely degenerate semiconductor must lie more than* 5 *kT below the top of the valence band.*

If it is remembered that the energy scale for holes is in the opposite direction to that for electrons it will follow that all results for holes should be analogous to the results for electrons provided the characteristics of electrons are replaced by those of holes.

Write the expressions for hole concentration for two limiting cases: non-degenerate semiconductor

$$p = N_v e^{-\frac{F-E_v}{kT}}; \quad N_v = 2\left(\frac{2\pi m_{pd}^* kT}{h^2}\right)^{3/2}, \qquad (29.34)$$

and completely degenerate semiconductor

$$p = \frac{8\pi}{3}\left(\frac{2m_{pd}^*}{h^2}\right)^{3/2}(E_v - F)^{3/2}. \qquad (29.35)$$

Summary of Secs. 28-29

1. The electron n and hole p concentrations are calculated with the aid of distribution function integrals over the Brillouin zone:

$$n = n(\mathbf{r}, t) = \frac{1}{4\pi^3}\int\limits_{(V_\mathbf{k})} f(\mathbf{r}, \mathbf{k}, t)\, d\tau_\mathbf{k}, \qquad (29.1\text{s})$$

$$p = p(\mathbf{r}, t) = \frac{1}{4\pi^3}\int\limits_{(V_\mathbf{k})} f_p(\mathbf{r}, \mathbf{k}, t)\, d\tau_\mathbf{k}. \qquad (29.2\text{s})$$

2. The Fermi-Dirac distribution function for electrons and holes depends on energy and temperature, therefore to calculate electron and hole concentrations one should instead of integrating over the Brillouin zone volume integrate over the energy inside an energy band. To this end the concept of states density $N(E)$ is introduced. This is the number of states per unit energy interval per unit crystal volume. $N(E)$ is defined by the condition

$$dS = N(E)\, dE = \frac{d\tau_\mathbf{k}}{8\pi^3}. \qquad (29.3\text{s})$$

The $N(E)$ function is determined by the shape of constant-energy surfaces.

3. In case of quadratic dependence of energy on quasimomentum we obtain for $N(E)$ near the minimum energy E_c

$$N(E)=2\left(\frac{2\pi m_d^*}{h^2}\right)^{3/2}(E-E_c)^{1/2}, \tag{29.4s}$$

and near the maximum E_v

$$N(E)=2\left(\frac{2\pi m_{pd}^*}{h^2}\right)^{3/2}(E_v-E)^{1/2}. \tag{29.5s}$$

4. The effective mass for the density of states m_d^* is related to the effective mass tensor $\mathbf{m}^*$ and to the number M of equivalent energy extrema

$$m_d^*=(M^2m_1m_2m_3)^{1/3}. \tag{29.6s}$$

If the extremum is in the centre of the Brillouin zone so that $M=1$, and constant-energy surfaces are spheres, then $m_d^*=m^*$.

5. The expressions for electron and hole concentrations are

$$n=\frac{2N_c}{\sqrt{\pi}}\Phi_{1/2}(\xi); \tag{29.7s}$$

$$p=\frac{2N_v}{\sqrt{\pi}}\Phi_{1/2}(\eta), \tag{29.8s}$$

where N_c and N_v are effective numbers of states in the conduction and valence bands, respectively:

$$N_c=2\left(\frac{2\pi m_d^*kT}{h^2}\right)^{3/2}=4.82\cdot10^{15}\left(\frac{m_d^*}{m}\right)^{3/2}T^{3/2}=$$
$$=2.5\cdot10^{19}\left(\frac{m_d^*}{m}\right)^{3/2}\left(\frac{T}{300}\right)^{3/2}, \tag{29.9s}$$

$$N_v=2\left(\frac{2\pi m_{pd}^*kT}{h^2}\right)^{3/2}=4.82\cdot10^{15}\left(\frac{m_{pd}^*}{m}\right)^{3/2}T^{3/2}=$$
$$=2.5\cdot10^{19}\left(\frac{m_{pd}^*}{m}\right)^{3/2}\left(\frac{T}{300}\right)^{3/2}. \tag{29.10s}$$

6. The expressions for electron and hole concentrations in a non-degenerate semiconductor are

$$n=N_ce^{-\frac{E_c-F}{kT}}, \tag{29.11s}$$

$$p=N_ve^{-\frac{F-E_v}{kT}}. \tag{29.12s}$$

The product of electron and hole concentrations in a non-degenerate semiconductor is independent of the Fermi level position:

$$np=N_cN_ve^{-\frac{\Delta E_0}{kT}}. \tag{29.13s}$$

A semiconductor is termed non-degenerate if its Fermi level lies no less than kT above the top of the valence band and below the bottom of the conduction band. In other words, in a non-degenerate semiconductor the Fermi level lies inside the forbidden band.

7. A semiconductor is termed completely degenerate if its Fermi level lies inside an energy band no less than $5kT$ away from the energy extremum. If the Fermi level lies no less than $5kT$ above E_c electron concentration will be independent of temperature

$$n=\frac{8\pi}{3}\left(\frac{2m_d^*}{h^2}\right)^{3/2}(F-E_c)^{3/2}, \qquad (29.14s)$$

and hole concentration will be given by the expression (28.12s). If the Fermi level lies less than $5kT$ below E_v hole concentration will be independent of temperature:

$$p=\frac{8\pi}{3}\left(\frac{2m_{pd}^*}{h^2}\right)^{3/2}(E_v-F)^{3/2}, \qquad (29.15s)$$

and electron concentration will be given by the expression (29.11s).

8. For degenerate semiconductors for all dispersion laws

$$n=\frac{1}{4\pi^3}\int\limits_{(V_k)} f\,d\tau_k=\frac{1}{4\pi^3}V_F, \qquad (29.16s)$$

where V_F is the Brillouin zone volume inside the Fermi surface. For spherical energy surfaces $V_F=\frac{4\pi}{3}k_F^3$

$$n=\frac{1}{3\pi^2}k_F^3, \qquad (29.17s)$$

the radius of the Fermi sphere being:

$$k_F=(3\pi^2 n)^{1/3}. \qquad (29.18s)$$

9. For the Kane dispersion law in case of small k it follows from (28.53) and (29.18s) that

$$n=\frac{1}{3\pi^2}\left(\frac{3}{2P^2}\right)^{3/2}F^{3/2}(E_g+F)^{3/2} \qquad (29.19s)$$

and

$$n=\frac{8\pi}{3}\left(\frac{2m_n^{0*}}{h^2}\right)^{3/2}F^{3/2}\left(1+\frac{F}{E_g}\right)^{3/2}. \qquad (29.20s)$$

30. ELECTRIC NEUTRALITY EQUATION

The expressions for n and p enable the electron and hole concentrations to be calculated provided the position of the Fermi level is known. However, the Fermi level itself depends on temperature and carrier concentration. Its position may change greatly when impurities responsible for localized states are introduced. It is quite natural because the Fermi level determines the distribution of electrons over their states. By introducing impurities we create localized states in the forbidden band which may be occupied both by electrons and holes. When discrete levels are created in the forbidden band the re-distribution of electrons over states is regulated by changes in the position of the Fermi level.

The equation used to calculate F, generally called the electric neutrality equation, has an obvious physical meaning. First of all suppose that the semiconductor is doped with dohor and acceptor impurities whose concentrations are N_d and N_a, respectively. As a result of thermal ionization a number of electrons and holes are created in the semiconductor. Free charge carriers are created as a result of the ionization both of impurity and matrix atoms; in other words, the semiconductor will contain free carriers and ions. *The total charge of all charged particles both in the crystal as a whole and in any physically small volume should be zero—this is the electroneutrality condition which is valid for an uncharged body.*

We will write the electric neutrality condition for a unit volume of the substance. To this end the number of positive and negative particles should be calculated.

The electrons are generated by the ionization both of donor impurity and matrix atoms. The transition of electrons from the valence band to the conduction band or to acceptor atoms results in the generation of free holes.

The negative charge is created by free electrons and acceptor ions. It is equal to $(n+N_a^-)e^-$. The positive charge equal to $(p+N_d^+)e^+$ is created by free holes and donor ions.

The electric neutrality condition may be written in the form

$$(p+N_d^+)e^+ + (n+N_a^-)e^- = 0. \tag{30.1}$$

Since $e^- = -e^+$ the electric neutrality equation will be

$$(n+N_a^-)-(p+N_d^+) = 0. \tag{30.2}$$

Denoting the number of electrons and holes occupying donor and acceptor levels by n_d, p_d, n_a, p_a, we may write some obvious equations

$$\begin{aligned} n_d &= N_d - N_d^+ = N_d - p_d; \quad N_d^+ = N_d - n_d = p_d, \\ p_a &= N_a - N_a^- = N_a - n_a; \quad N_a^- = N_a - p_a = n_a. \end{aligned} \tag{30.3}$$

Now the electric neutrality equation (30.1) or (30.2) may be written in one of these forms:

$$(n+n_a)-(p+p_d) = 0, \tag{30.4}$$

$$n + n_d - p - p_a = N_d - N_a. \tag{30.5}$$

To derive an equation for the calculation of the Fermi level position the quantities in the electric neutrality equation (30.5) should be expressed in terms of the Fermi levels. The expression for n and p being already known, it remains to find n_d and p_d (or n_a and p_a). To find the number of electrons at an impurity level one should know the distribution function of electrons over impurity states.

Evidently, the Fermi-Dirac function may not be directly used to describe the distribution of electrons over impurity states since it is valid for the case when there may be two electrons with opposing spins in one state. However, in each localized impurity state (with energy E_d or E_a) there may not be more than one electron. If a second electron is placed in such a state the original energy of the state E_d or E_a will change greatly because of strong electrostatic interaction between the two electrons. In other words, energy levels of singly and doubly ionized impurity atoms are different. The electron distribution function should reflect the impossibility of two electrons occupying one localized state. *The resulting distribution function for impurity states* obtained with the aid of the Gibbs method for a system with a variable number of particles *is of the form*

$$f=\frac{1}{\frac{1}{g_i}e^{\frac{E_i-F}{kT}}+1}, \qquad (30.6)$$

where g_i is the degeneracy factor for the ith impurity state. For $E_i=E_d$, which corresponds to a donor impurity, $g_i=2$. For $E_i=$ $=E_a$ (acceptor impurity) $g_i=\frac{1}{2}$. Hence, *the distribution of electrons over donor states is given by the expression*

$$f=\frac{1}{\frac{1}{2}\cdot e^{\frac{E_d-F}{kT}}+1}. \qquad (30.7)$$

The distribution of electrons over acceptor states is of the form

$$f=\frac{1}{2e^{\frac{E_a-F}{kT}}+1}. \qquad (30.8)$$

Accordingly, *the distribution function for holes is of the form*

$$f_p=\frac{1}{2e^{\frac{F-E_d}{kT}}+1}; \quad f_p=\frac{1}{\frac{1}{2}e^{\frac{F-E_a}{kT}}+1}. \qquad (30.9)$$

Now the number of electrons and holes occupying impurity levels may be easily found. Find, for example, n_d:

$$n_d=\int N_d(E)\,f\,dE=N_d\int \delta(E-E_d)\frac{1}{\frac{1}{2}e^{\frac{E-F}{kT}}+1}\,dE=\frac{N_d}{\frac{1}{2}e^{\frac{E_d-F}{kT}}+1}. \tag{30.10}$$

In the same way one may find n_a, p_a, p_d:

$$n_a=\frac{N_a}{2e^{\frac{E_a-F}{kT}}+1}, \tag{30.11}$$

$$p_a=\frac{N_a}{\frac{1}{2}e^{\frac{F-E_a}{kT}}+1}; \quad p_d=\frac{N_d}{2e^{\frac{F-E_d}{kT}}+1}. \tag{30.12}$$

Write the electric neutrality equation (30.5) taking into account the explicit expressions (29.1), (29.13), (30.10) and (30.12) for n, p, n_d and p_a:

$$4\pi\left(\frac{2m_d^*}{h^2}\right)^{3/2}\int_{E_c}^{\infty}\frac{(E-E_c)^{1/2}\,dE}{e^{\frac{E-F}{kT}}+1}+\frac{N_d}{\frac{1}{2}e^{\frac{E_d-F}{kT}}+1}-$$

$$-4\pi\left(\frac{2m_{pd}^*}{h^2}\right)^{3/2}\int_{-\infty}^{E_v}\frac{(E_v-E)^{1/2}\,dE}{e^{\frac{F-E}{kT}}+1}-\frac{N_a}{\frac{1}{2}e^{\frac{F-E_a}{kT}}+1}=N_d-N_a. \tag{30.13}$$

This equation considered as one for the Fermi level may be reduced to an algebraic fourth order equation. Its general solution is very complicated, and we, therefore, will discuss some specific cases of practical importance. Besides it should be kept in mind that the equation is written for the case of only one donor or one acceptor level. Otherwise the equation (30.13) will be still more complicated.

31. INTRINSIC SEMICONDUCTOR

A semiconductor without impurities ($N_d=N_a=0$) *is termed pure or intrinsic. Its electric neutrality equation is of the form*

$$n-p=0, \quad \text{or} \quad n=p, \tag{31.1}$$

i.e. the transition of an electron from the valence band results in the creation of a hole in it, therefore *the numbers of electrons and* *holes are equal* (Fig. 49). The equation for calculating the position of the Fermi level (30.13) in condition (31.1) is as follows:

$$\frac{2N_c}{\sqrt{\pi}}\Phi_{1/2}(\xi)=\frac{2N_v}{\sqrt{\pi}}\Phi_{1/2}(\eta). \qquad (31.2)$$

Suppose $N_v=N_c$ which entails the equality of effective masses for states densities $m_d^*=m_{pd}^*$. In this case

$$\Phi_{1/2}(\xi)=\Phi_{1/2}(\eta) \qquad (31.3)$$

and, therefore, $\xi=\eta$. Substituting the expressions for ξ and η we obtain

$$\frac{F-E_c}{kT}=\frac{E_v-F}{kT}, \qquad (31.4)$$

whence

$$F=\frac{E_v+E_c}{2}=E^i, \qquad (31.5)$$

i.e. *the Fermi level is independent of temperature and lies exactly in the middle of the forbidden band (coincides with E^i).*

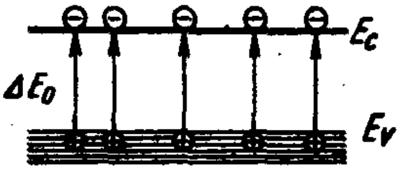


Fig. 49. The generation of conduction electrons and holes in an intrinsic semiconductor

The forbidden band width for semiconductors is much greater than $2kT$; for this reason $\xi\ll-1$ and $\eta\ll-1$; the intrinsic semiconductor is non-degenerate, the Fermi integral is equal to $\frac{\sqrt{\pi}}{2}e^{\xi}$ and $\frac{\sqrt{\pi}}{2}e^{\eta}$, respectively, and we obtain the following expressions for the concentrations of electrons and holes:

$$n=N_c e^{\xi}=N_c e^{\frac{E_c-F}{kT}}=N_c e^{-\frac{\Delta E_0}{2kT}}=N_c e^{-\frac{\delta E_a}{kT}} \qquad (31.6)$$

$$p=N_v e^{\eta}=N_v e^{-\frac{F-E_v}{kT}}=N_v e^{-\frac{\Delta E_0}{2kT}}=N_v e^{-\frac{\delta E_a}{kT}}. \qquad (31.7)$$

The activation energy δE_a for an intrinsic semiconductor is equal to half the forbidden band width.

We obtain the expressions (31.6) and (31.7) in the assumption of $N_v=N_c$. Suppose $N_v\neq N_c$. Since, however, they are of the same order of magnitude, the Fermi integral may, as before, be represented in the form $\frac{\sqrt{\pi}}{2}e^{\xi}$ and $\frac{\sqrt{\pi}}{2}e^{\eta}$, respectively; cancelling out $\frac{\sqrt{\pi}}{2}$ in (31.2) we obtain

$$n=N_c e^{\xi}=N_v e^{\eta}=p, \tag{31.8}$$

or

$$\xi-\eta=\ln\frac{N_v}{N_c}=\frac{F-E_c}{kT}-\frac{E_v-F}{kT}. \tag{31.9}$$

It follows from (31.9) that

$$F=\frac{E_c+E_v}{2}+\frac{kT}{2}\ln\frac{N_v}{N_c}=\frac{E_c+E_v}{2}-kT\ln\left(\frac{N_c}{N_v}\right)^{1/2}. \tag{31.10}$$

Since

$$\frac{N_c}{N_v}=\left(\frac{m_d^*}{m_{pd}^*}\right)^{\frac{3}{2}}=\left(\frac{m_{nd}^*}{m_{pd}^*}\right)^{\frac{3}{2}},$$

then

$$F=\frac{E_c+E_v}{2}+kT\ln\left(\frac{m_{pd}^*}{m_{nd}^*}\right)^{3/4}. \tag{31.11}$$

m_{nd}^* denotes the quantity which was formerly denoted by m_d^*. We see from here that *the Fermi level at $T=0$ lies in the middle of the forbidden band, and that its position depends linearly on temperature* (Fig. 50). *As the temperature rises the Fermi level draws nearer to the band the density of states in which is less and which, therefore, will be filled sooner*. Although there are no electrons at the Fermi level in an intrinsic semiconductor the definition that when $E=F$ $f=1/2$ remains true. *The concentration of intrinsic charge carriers is determined from the equation*

$$n_i=n=p=\sqrt{N_cN_v}\,e^{-\frac{\Delta E_0}{2kT}}. \tag{31.12}$$

Substituting numerical values of constants involved in the expressions for N_c and N_v we obtain for the concentration n_i

$$n_i=4.82\cdot 10^{15}\cdot\left(\frac{m_{nd}^*m_{pd}^*}{m^2}\right)^{3/4}T^{3/2}e^{-\frac{\Delta E_0}{2kT}}, \tag{31.13}$$

and

$$np=n_i^2=2.31\cdot 10^{31}\left(\frac{m_{nd}^*m_{pd}^*}{m^2}\right)^{3/2}T^3e^{-\frac{\Delta E_0}{kT}}. \tag{31.14}$$

The graphical dependence of $\ln n_i$ on inverse temperature will practically be a straight line:

$$\ln n_i=\text{const}-\frac{3}{2}\ln\frac{1}{T}-\frac{\Delta E_0}{2k}\cdot\frac{1}{T}, \tag{31.15}$$

since the term $\ln\frac{1}{T}$ will be negligible as compared to the linear one. The slope of the line is determined by the forbidden band width

$$\tan\varphi = -\frac{\Delta E_0}{2k},$$

whence

$$\Delta E_0 = 2k\,|\tan\varphi|, \qquad (31.16)$$

tan φ may be measured from the $\left(\ln n_i \text{ vs } \frac{1}{T}\right)$ graph (Fig. 51).

Assess the intrinsic carrier concentration in some semiconductors. For germanium and silicon $\left(\frac{m^*_{nd}m^*_{pd}}{m^2}\right)^{3/2}$ is equal to 0.299 and 0.719, respectively, and the corrsponding approximate values of n_i at $T \cong 300$ °K are 2×10^{13} cm^{-3} and $2\cdot10^{10}$ cm^{-3}.

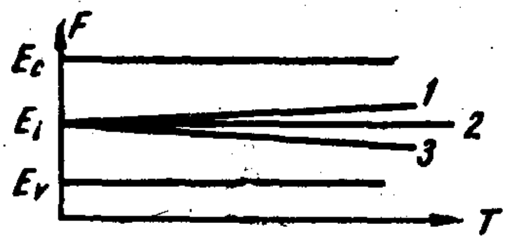


Fig. 50. The temperature dependence of the Fermi level in an intrinsic semiconductor:

1—$m^*_{nd} < m^*_{pd}$, 2—$m^*_{nd} = m^*_{pd}$; 3—$m^*_{nd} > m^*_{pd}$;

$F = E^i$; $E^i = \frac{E_v + E_c}{2}$ —the middle of the forbidden band

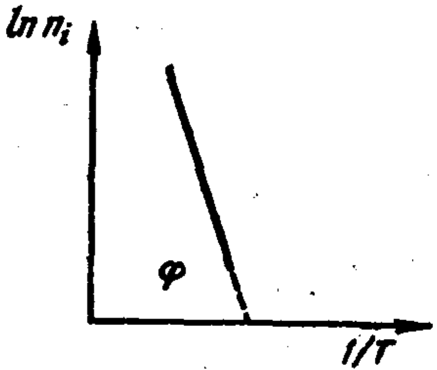


Fig. 51. The dependence of $\ln n_i$ on the inverse temperature in an intrinsic semiconductor

In compliance with the expression for n_i the carrier concentration turns zero at $T = 0$, and the resistance of an intrinsic semiconductor should at $T \to 0$ tend to infinity. Real semiconductors, however, always contain some residual impurities which provide for finite conductivities at all temperatures.

Now we will discuss the problem of carrier concentration in semiconductors with degenerate bands, such as p-germanium and p-silicon.

Each band is characterized by its own effective mass m_{p1}^* and m_{p2}^*. The density of states in them may be written in the form

$$N_1(E) = 2\pi\left(\frac{2m_{pd1}^*}{h^2}\right)^{3/2}(E_v - E)^{1/2}, \tag{31.17}$$

$$N_2(E) = 2\pi\left(\frac{2m_{pd2}^*}{h^2}\right)^{3/2}(E_v - E)^{1/2}. \tag{31.18}$$

We may introduce the total density of states

$$N(E) = N_1(E) + N_2(E) = 2\pi\left(\frac{2m_{pd}^*}{h^2}\right)^{3/2}(E_v - E)^{1/2} \tag{31.19}$$

where

$$(m_{pd}^*)^{3/2} = (m_{pd1}^*)^{3/2} + (m_{pd2}^*)^{3/2}. \tag{31.20}$$

Subsequently we may use the expression (31.13) for the carrier concentration. For a g-fold degeneracy one should add up the effective masses for the densities of states taking account of the number of maxima M (in the power of 3/2) for holes of each type (the same, evidently, applies to electrons). It is easily seen that *the ratios of concentrations of holes of different types is determined solely by their densities of states*:

$$\frac{p_1}{p_2} = \frac{N_{v1}}{N_{v2}} = \left(\frac{m_{pd1}^*}{m_{pd2}^*}\right)^{3/2}, \tag{31.21}$$

i.e. *there should be less light carriers than heavy ones*. For germanium and silicon the ratio of light to heavy hole masses $\frac{m_l^*}{m_h^*}$ is 0.13 and 0.3, respectively, therefore the concentrations of light holes should be around 0.047 and 0.165 of the heavy hole concentrations. The concentration of light holes amounts in per cent to the total hole concentration to 4.5% in germanium and 14% in silicon.

In conclusion of the section we will discuss the case of a narrow band: $\Delta E_0 \ll kT$. If in a substance the overlapping of the conduction and the valence band is Δ, such a substance is termed semimetal. Here we will consider a pure, or intrinsic, semimetal with a negative forbidden band:

$$E_c - E_v = -\Delta. \tag{31.22}$$

When $kT \ll \Delta$ the semimetal may be degenerate both with electrons and holes, and for this reason the expressions (29.14s) and (29.15s) may be used to yield electron and hole concentrations

$$n = \frac{8\pi}{3}\left(\frac{2m_{nd}^*}{h^2}\right)^{3/2}\delta_n^{3/2}; \tag{31.23}$$

$$p = \frac{8\pi}{3}\left(\frac{2m_{pd}^*}{h^2}\right)^{3/2}\delta_p^{3/2}, \tag{31.24}$$

where

$$\delta_n = F - E_c; \quad \delta_p = E_v - F; \quad \delta_n + \delta_p = E_v - E_c = \Delta. \tag{31.25}$$

With the increase in temperature the degeneracy of one of the gases (electron or hole) may vanish. This is more probable in case of the hole gas because the effective mass of holes is greater than that of electrons, and the conduction band will be filled sooner. When the Fermi level leaves the overlapping strip and rises above the top of the valence band, only the electron gas degeneracy will remain.

In the double degeneracy region we obtain from (31.23-24) and the condition $n = p$

$$m^*_{nd}\delta_n = m^*_{pd}\delta_p \tag{31.26}$$

and

$$\delta_n = \frac{m^*_{pd}}{m^*_{nd} + m^*_{pd}}\Delta; \quad \delta_p = \frac{m^*_{nd}}{m^*_{nd} + m^*_{pd}}\Delta. \tag{31.27}$$

The intrinsic carrier concentration in a semimetal in the region of electron and hole degeneracy is

$$n = p = n_i = \frac{8\pi}{3}\left(\frac{2}{h^2} : \frac{m^*_{nd}\,m^*_{pd}}{m^*_{nd} + m^*_{pd}}\right)^{3/2}\Delta^{3/2}. \tag{31.28}$$

It follows from (31.28) that intrinsic carrier concentration in a semimetal in a degenerate state is determined by the energy band overlapping Δ for $\Delta = 0$, $n_i = 0$.

Simultaneous electron and hole degeneracy is impossible in a semiconductor with a positive forbidden band. However, if the effective mass of electrons is much less than that of holes the Fermi level may turn out to be inside the conduction band as a result of which the electron gas in an intrinsic narrow-band semiconductor may turn out to be degenerate. Such is the situation, for example, in quicksilver telluride in which the thermal forbidden band width (but not the Kane parameter E_g) is zero.

32. EXTRINSIC SEMICONDUCTOR. IMPURITY OF ONE TYPE

Consider a semiconductor doped with impurity of one type, donors, for example, so that $N_d \neq 0$ and $N_a = 0$. The electric neutrality equation (30.5) assumes the form

$$n + n_d - p = N_d \tag{32.1}$$

or

$$n = p + N_d^+. \tag{32.2}$$

The meaning of equation (32.2) is quite obvious. *Free electrons are generated by electron transitions from the valence band which* *results in the generation of p free holes, and from the impurity level, which results in the generation of N_d^+ donor impurity ions* (Fig. 52). Equation (32.2) is a third degree equation with respect to F. There are, however, some comparatively simple cases when it is easy to determine the position of the Fermi level and thereby the electron and hole concentrations.

Note first of all that in a non-degenerate semiconductor it is sufficient to determine the concentration of charge carriers of only one sign leaving the concentration of carriers of another sign to be calculated from the relations (30.5) and (31.12):

$$np = n_i^2 = N_c N_v e^{-\frac{\Delta E_0}{kT}}. \qquad (32.3)$$

It follows from the expression (32.3) that *in a non-degenerate semiconductor the product of electron and hole concentrations is independent of the Fermi level position and, hence, of the doping of the semiconductor and is equal to the square of the concentration of one type of carriers in the intrinsic semiconductor*. The relation (32.3) enables the concentration of carriers of one type to be found with the aid of that of the other. For instance, in case electron concentration is known, hole concentration may be found with the aid of the relation

$$p = \frac{n_i^2}{n} = \frac{N_c N_v e^{-\frac{\Delta E_0}{kT}}}{n}. \qquad (32.4)$$

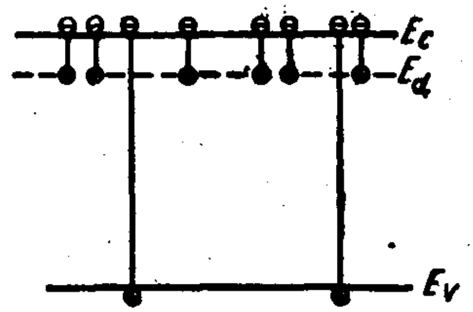


Fig. 52. Thermal generation of charge carriers in a donor-doped semiconductor

Turn again to equation (32.2). Free electrons are generated as a result of the ionization both of the matrix and of the impurity atoms, the part played by each of these processes being different at different temperatures. To transfer an electron from the valence to the conduction band energy should be expended equal to the forbidden band width ΔE_0. At the same time energy needed to transfer an electron from an impurity level to the conduction band is equal to the impurity atom ionization energy ΔE_d which is much less than the forbidden band width. Therefore *at low temperatures the main part will be played by electron transitions from the impurity level* and, hence, $p \ll N_d^+$. At some temperature practically all the impurity atoms will be ionized, and further increase in electron concentration n with rising temperature will be due to the ionization of the matrix atoms. *At high temperatures $p \gg N_d^+ = N_d$, and the semiconductor becomes intrinsic.* The definition of high or low temperature must be related to impurity concentration. *Any temperature may at the same time be low or high depending on the* *impurity concentration: it may be high for low concentrations and low — for high concentrations.*

1. **Low temperatures.** Consider the case of low temperatures when impurity concentration plays the leading part. In this case $p \ll N_d^+$, and the equation (32.2) is simplified:

$$n = N_d^+, \quad \text{or} \quad n = p_d. \tag{32.5}$$

Substituting the expressions for n and p_d into (32.5) we obtain

$$n = N_c e^{-\frac{E_c - F}{kT}} = p_d = \frac{N_d}{2e^{\frac{F - E_d}{kT}} + 1}. \tag{32.6}$$

Denoting $e^{\frac{F}{kT}} = x$ we obtain

$$N_c e^{-\frac{E_c}{kT}} x \left(1 + 2e^{-\frac{E_d}{kT}} x\right) = N_d, \tag{32.7}$$

or,

$$2N_c e^{-\frac{E_c + E_d}{kT}} x^2 + N_c e^{-\frac{E_c}{kT}} x - N_d = 0, \tag{32.8}$$

$$x^2 + \frac{1}{2} e^{\frac{E_d}{kT}} x - \frac{N_d}{2N_c} e^{\frac{E_c + E_d}{kT}} = 0. \tag{32.9}$$

Solving the equation (32.9) for x we obtain

$$x = \frac{1}{4} e^{\frac{E_d}{kT}} \left(\pm \sqrt{1 + \frac{8N_d}{N_c} e^{\frac{\Delta E_d}{kT}}} - 1\right). \tag{32.10}$$

Since $x > 0$, minus before the radical should be omitted; therefore, from (32.10) we obtain for F:

$$F = kT \ln \left\{ \frac{1}{4} e^{\frac{E_d}{kT}} \left(\sqrt{1 + \frac{8N_d}{N_c} e^{\frac{\Delta E_d}{kT}}} - 1 \right) \right\}. \tag{32.11}$$

We would like to repeat that (32.11) holds for such low temperatures that $p \ll N_d^+$ and $n = N_d^+$, i.e. *conduction electrons are generated mainly as the result of donor impurity ionization.*

Consider two limiting cases for F. As temperature rises $e^{\frac{\Delta E_d}{kT}}$ tends to unity, N_c increases and may exceed N_d. However, for sufficiently low temperatures the inequality

$$\frac{8N_d}{N_c} e^{\frac{\Delta E_d}{kT}} \gg 1 \tag{32.12}$$

may hold. Therefore we obtain for x

$$x=\frac{1}{4}e^{\frac{E_d}{kT}}\sqrt{\frac{8N_d}{N_c}}e^{\frac{E_c-E_d}{2kT}}=\sqrt{\frac{N_d}{2N_c}}e^{\frac{E_c+E_d}{2kT}}=e^{\frac{F}{kT}} \tag{32.13}$$

Write the expression for the position of the Fermi level using (32.13):

$$F=\frac{E_c+E_d}{2}+\frac{kT}{2}\ln\frac{N_d}{2N_c}. \tag{32.14}$$

At $T=0$

$$F=\frac{E_c+F_d}{2}, \tag{32.15}$$

i.e. *the Fermi level lies midway between the bottom of the conduction band and the impurity level. The Fermi level rises as the temperature is increased, reaches a maximum at some temperature and then drops again.* When $2N_c=N_d$ it is again midway between E_c and E_d. The validity of the formula (32.14) in this temperature range is, however, conditional upon the concentration N_d; for small impurity concentrations N_c equals $\frac{N_d}{2}$ at such temperatures that $e^{\frac{\Delta E_d}{2kT}}\gg 1$. If, on the other hand, N_d is large so that $e^{\frac{\Delta E_d}{2kT}}\cong 1$ the formula (32.14) for F will not hold.

Find the electron concentration

$$n=N_ce^{-\frac{E_c-F}{kT}}=\sqrt{\frac{N_cN_d}{2}}e^{-\frac{\Delta E_d}{2kT}}. \tag{32.16}$$

It is remarkable that *at low temperatures the electron concentration depends on the impurity concentration in the power of* 1/2. The graph $\left(\ln n, \frac{1}{T}\right)$ is almost a straight line with the slope which is equal to $\frac{\Delta E_d}{2k}$ while for the intrinsic carrier concentration line the slope is determined by $\frac{\Delta E_0}{2k}$.

Consider now the opposite case:

$$\frac{8N_d}{N_c}e^{\frac{\Delta E_d}{kT}}\ll 1; \quad \text{or} \quad N_c\gg 8N_d. \tag{32.17}$$

For the condition $8N_d\ll N_c$ to be realized a sufficiently high temperature is essential.

Expanding the radical (32.10) into a series and taking only the first term we obtain

$$x=\frac{1}{4}e^{\frac{E_d}{kT}}\left(1+\frac{4N_d}{N_c}e^{\frac{\Delta E_d}{kT}}+\ldots-1\right)=\frac{N_d}{N_c}e^{\frac{E_c}{kT}}. \tag{32.18}$$

The expression for the Fermi level that follows from (32.18) is

$$F = E_c + kT \ln \frac{N_d}{N_c}. \tag{32.19}$$

Since it is valid for $N_c \gg N_d$ the logarithm in (32.19) should be negative, and the Fermi level should sink with the rise in temperature. Find the electron concentration for this case:

$$n = N_c e^{-\frac{E_c - F}{kT}} = N_c e^{\ln \frac{N_d}{N_c}} = N_d, \tag{32.20}$$

i.e. *the electron concentration is independent of temperature and equal to the impurity concentration. This temperature range is termed impurity depletion range.* Recall that *the charge carriers are termed majority carriers when their concentration exceeds that of intrinsic carriers n_i at the given temperature. When their concentration is below n_i they are termed minority carriers*. Thus, *in donor-doped semiconductors electrons are majority carriers.* In the same way we may say that *in the impurity depletion range the majority carrier concentration remains constant; the concentration of minority carriers, on the other hand, should change rapidly with temperature.* Indeed, we may write from (32.4)

$$p = \frac{n_i^2}{n} = \frac{n_i^2}{N_d} = \frac{N_c N_v}{N_d} e^{-\frac{\Delta E_0}{kT}} \tag{32.21}$$

This expression will be valid as long as hole concentration remains much less than the electron concentration:

$$p \ll n = N_d^+ = N_d. \tag{32.22}$$

2. **High temperatures.** As temperature rises the concentration of holes increases and may become comparable to the electron concentration. When this is the case the equation (32.5) is to be replaced by the general equation (32.2) which may be substantially simplified. Indeed, write (32.2) taking into account (32.20):

$$n = p + N_d \tag{32.23}$$

It is valid for the case when *all impurity is ionized, and the ionization of the matrix has to be taken into account.* Write the equation (32.23) for a non-degenerate semiconductor:

$$n = \frac{n_i^2}{n} + N_d, \tag{32.24}$$

$$n^2 - nN_d - n_i^2 = 0. \tag{32.25}$$

Solving this equation we obtain

$$n=\frac{N_d}{2}\left(1\pm\sqrt{1+\frac{4n_i^2}{N_d^2}}\right). \tag{32.26}$$

Since the expression under the radical sign exceeds unity, and $n>0$, the minus sign in front of the radical should be omitted. Write now the expression for the electron and hole concentrations:

$$n=\frac{N_d}{2}\left(1+\sqrt{\frac{4n_i^2}{N_d^2}}\right), \tag{32.27}$$

$$p=\frac{2n_i^2}{N_d\left(1+\sqrt{1+\frac{4n_i^2}{N_d^2}}\right)}. \tag{32.28}$$

Taking into account the relation between n and F we obtain from (32.27)

$$F=E_c+kT\ln\left\{\frac{N_d}{2N_c}\left(1+\sqrt{1+\frac{4n_i^2}{N_d^2}}\right)\right\}=$$

$$=E_c+kT\ln\left\{\frac{N_d}{2N_c}\left(1+\sqrt{1+\frac{4N_cN_v}{N_d^2}e^{-\frac{\Delta E_0}{kT}}}\right)\right\}. \tag{32.29}$$

Consider two extreme cases. If

$$\frac{4n_i^2}{N_d^2}\ll 1, \tag{32.30}$$

it follows from (32.27) and (32:29) that

$$n=N_d;\ p=\frac{n_i^2}{N_d};\ F=E_c+kT\ln\frac{N_d}{N_c}, \tag{32.31}$$

which is in full accord with results obtained previously for the depletion temperature range. If, on the other hand,

$$\frac{4n_i^2}{N_d^2}\gg 1, \tag{32.32}$$

then

$$n=p=n_i;\ F=\frac{E_c+E_v}{2}+\frac{kT}{2}\ln\frac{N_v}{N_c}, \tag{32.33}$$

as it should be for an intrinsic semiconductor.

Thus, *in the entire temperature range there are two expressions to describe the position of the Fermi level in a non-degenerate* *semiconductor. The expression*

$$F = E_d + kT \ln\left\{\frac{1}{4}\left(\sqrt{1 + \frac{8N_d}{N_c}e^{\frac{\Delta E_d}{kT}} - 1}\right)\right\} \tag{32.34}$$

is valid from $T = 0$ up to the depletion temperature T_{dep}. For temperatures in excess of T_{dep} the expression

$$F = E_c + kT \ln\left\{\frac{N_d}{2N_c}\left(1 + \sqrt{1 + \frac{4n_i^2}{N_d^2}}\right)\right\} \tag{32.35}$$

holds.

The expressions (32.34) and (32.35) describe the temperature dependence of the Fermi level in a specific semiconductor. At $T = 0$ the Fermi level lies midway between the bottom of the conduction band and the donor level. As temperature rises the Fermi level rises up to E_c, then, because of an increase in N_c, it goes through a maximum and begins to sink.

The electron concentration increases in the process owing to impurity ionization. At some temperature F will be equal to E_d, and there will be $\frac{2}{3}N_d$ electrons at the impurity level, and $n = \frac{N_d}{3}$ in the conduction band. As the Fermi level sinks lower the semiconductor goes over to the depletion range: here the impurity is completely ionized, electron concentration remains constant, hole concentration increases, and the Fermi level approaches the middle of the forbidden band. As the Fermi level approaches the middle of the forbidden band the hole concentration increases, while electron concentration remains practically the same. A further increase in hole concentration with the rise in temperature is accompanied by an increase in electron concentration until the equality $n = p$ is reached, and the semiconductor changes from the impurity to the intrinsic type.

The transition temperature from impurity to intrinsic conductivity depends on the concentration of impurity in the specific semiconductor and on the forbidden band width for fixed impurity concentration. If the transition from impurity to intrinsic conductivity is defined symbolically by the condition $p = N_d$, or $n = 2N_d$, the transition temperature, as may easily be seen, will be determined by the equation:

$$pn = n_i^2 = 2N_d^2 = N_c N_v e^{-\frac{\Delta E_0}{kT_{dep}}}, \tag{32.36}$$

$$T_{dep} = \frac{\Delta E_0}{k \ln \frac{N_c(T_{dep})\, N_v(T_{dep})}{2N_d^2}}.$$

For a fixed N_d the temperature of transition to intrinsic conductivity is the higher the higher is ΔE_0. For a given semiconductor the transition temperature is the higher the higher is its impurity concentration.

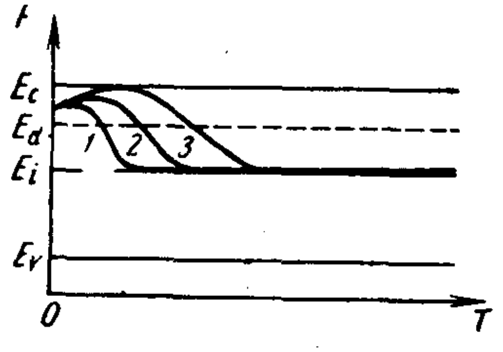


Fig. 53. The temperature dependence of the Fermi level in a donor-doped semiconductor:

$1-N_{d1}$; $2-N_{d2}$;
$3-N_{d3}$ ($N_{d1}<N_{d2}<N_{d3}$)

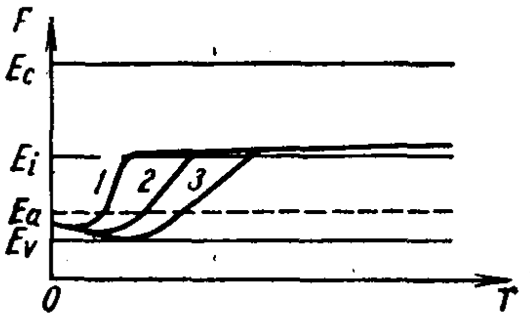


Fig. 54. The temperature dependence of the Fermi level in an acceptor-doped semiconductor:

$1-N_{a1}$; $2-N_{a2}$; $3-N_{a3}$;
($N_{a1}<N_{a2}<N_{a3}$)

Figure 53 shows the dependence of the Fermi level position on temperature for three different impurity concentrations.

Write the expression for F and p for the case of acceptor impurity:

$$p=n+p_a=n+N_a^-. \tag{32.37}$$

Solving the equation (32.37) in the same way as was done in the case of donor impurity we obtain for $n\ll N_a^-$

$$F=E_a-kT\ln\left\{\frac{1}{4}\left(\sqrt{1+\frac{8N_c}{N_v}e^{\frac{\Delta E_a}{kT}}}-1\right)\right\}. \tag{32.38}$$

In case $N_a^-=N_a$, and $n\gg N_a$ we obtain likewise

$$F=E_v-kT\ln\left\{\frac{N_a}{2N_v}\left(1+\sqrt{1+\frac{4n_i^2}{N_a^2}}\right)\right\};$$

$$p=\frac{N_a}{2}\left(1+\sqrt{1+\frac{4n_i^2}{N_a^2}}\right). \tag{32.39}$$

Figure 54 shows the temperature dependence of the Fermi level in a hole-type semiconductor.

To conclude the section we will assess the temperature at which the impurity is depleted. Since in the depletion range both equations (32.17) (for the lower limit) and (32.30) (for the upper limit) should be satisfied, uniting them we obtain

$$\frac{N_c(T_l)}{8}\gg N_d>2n_i(T_{ap}). \tag{32.40}$$

Find the lower limit of the depletion range by making it satisfy the condition

$$N_c(T_l) = N_d. \tag{32.41}$$

Taking into accout (28.9s) we obtain

$$\left(\frac{T_l}{300}\right)^{3/2} = \left(\frac{m}{m_d^*}\right)^{3/2} \cdot 4 \cdot 10^{-18} N_d, \tag{32.42}$$

$$\frac{T_l}{300} = \frac{m}{m_d^*} 2.5 \left(\frac{N_d}{10^{18}}\right)^{2/3}. \tag{32.43}$$

Let $m_d^* = 0.25m$, in this case

$$\frac{T_l}{300} = 10 \left(\frac{N_d}{10^{18}}\right)^{2/3}. \tag{32.44}$$

For $N_d = 10^{18}$ cm^{-3} $T_l = 3000$ °K, for $N_d = 10^{15}$ cm^{-3} $T_l = 30$ °K, i.e. in the first instance depletion sets in only at 3000 °K, and in the second—already at 30 °K.

The upper limit may be assessed with the aid of relation (32.36). It must, however, be taken into account that T_{ap} enters N_c and N_v.

To assess the parts played by particular terms in the expression for n, p or F the value of kT should be compared with ΔE_0 or ΔE_d. We would like to remind that the Boltzmann constant $k = 1.38 \times \times 10^{-23}$ J·K^{-1} = 1.38×10^{-16} erg·K^{-1} = 8.6167×10^{-5} eV·K^{-1}. The reciprocal $\frac{1}{k} = 11\,605.4$ K·eV^{-1}, in other words, 1 eV corresponds to 11 605.4 °K. Table 8 shows the values of kT for some temperatures.

Table 8

T, °K	1	4.2	20	100	200	273	290	300	500
kT, eV	8.6×10^{-5}	3.6×10^{-4}	1.7×10^{-3}	8.6×10^{-3}	0.017	0.0235	0.0250	0.0258	0.043

The energy of ionization of impurity atoms, one should remember, is of the order of hundredths of an electron-volt.

Summary of Secs. 30-32

1. The Fermi energy is the Gibbs thermodynamic potential per particle. The other name for it is chemical potential. It is equal to the increase in the energy of a system of particles when one particle is added to it. For this reason the Fermi energy is determined by the total number of particles. In semiconductors the

Fermi energy is determined by the distribution of electrons over the energy levels of the valence band, the conduction band and discrete levels of localized states. The equation which determines the electron distribution over states is usually termed electric neutrality equation

$$n+n_d-p-p_a=N_d-N_a. \tag{32.1s}$$

If n, n_d, p and p_a are expressed in terms of F the result will be the equation (30.13) which generally enables the Fermi level to be found.

2. Electron distribution function for discrete states differs from electron distribution function for band states in that a localized state can accommodate not more than one electron:

$$f(E_t, T)=\frac{1}{\frac{1}{g}e^{\frac{E_t-E}{kT}}+1}. \tag{32.2s}$$

For the donor level $g=2$, and for the acceptor level $g=\frac{1}{2}$. $f_p(E_t, T)$ for holes is found from the condition

$$f_p(E_t, T)=1-f(E_t, T).$$

3. A semiconductor is termed intrinsic if it does not contain active impurities: $N_a=N_d=0$. The electric neutrality equation for an intrinsic semiconductor is of the form

$$n=p. \tag{32.3s}$$

This leads to the following expression for F:

$$F=\frac{E_v+E_c}{2}+\frac{kT}{2}\ln\frac{N_v}{N_c}=\frac{E_c+E_v}{2}+\frac{3kT}{4}\ln\frac{m^*_{pd}}{m^*_{nd}}. \tag{82.4s}$$

At $T=0$ the Fermi level lies in the middle of the forbidden band. As temperature increases it moves towards the band in which the effective mass for the density of states is less.

4. In an intrinsic semiconductor the charge carrier concentration is determined by the forbidden band width and temperature

$$n_i=n=p=\sqrt{N_cN_v}\,e^{-\frac{\Delta E_0}{2kT}}. \tag{32.5s}$$

5. In a semiconductor doped with impurity of one kind (donor, for example) the electric neutrality equation is of the form

$$n=p+N_d^+. \tag{32.6s}$$

At low temperatures, when $p\ll n$, the conductivity of the semiconductor is of the impurity type, and the position of the Fermi level is determined by the equation

$$F=E_d+kT\ln\left\{\frac{1}{4}\left(\sqrt{1+\frac{8N_d}{N_c}e^{\frac{\Delta E_d}{kT}}}-1\right)\right\}. \tag{32.7s}$$

At $T=0$ $F=\frac{E_c+E_d}{2}$ as temperature increases, F approaches the conduction band, passes through a maximum and begins to sink (see Fig. 53). The sinking rate is the highest when the impurity is fully ionized:

$$F=E_c+kT\ln\frac{N_d}{N_c}, \tag{32.8s}$$

$$n=N_d\,(p\ll n). \tag{32.9s}$$

This temperature range is termed impurity depletion (sometimes saturation) range.

6. At such temperatures that the impurity is fully ionized and the ionization of matrix atoms takes place the equation (32.6s) assumes the form:

$$n=N_d+p, \tag{32.10s}$$

and this leads to the following expression for F:

$$F=E_c+kT\ln\left\{\frac{N_d}{2N_c}\left(1+\sqrt{1+\frac{4n_i^2}{N_d^2}}\right)\right\}, \tag{32.11s}$$

which holds for regions from depletion range to intrinsic concentration.

7. The temperature of transition to intrinsic conductivity is the higher, the wider is the forbidden band and the higher is the impurity concentration.

33. SEMICONDUCTOR DOPED WITH BOTH ACCEPTOR AND DONOR IMPURITIES

Consider now the case of a semiconductor doped both with donor and acceptor impurities. Let $T=0$. At this temperature the electron system occupies the lowest energy states. The conduction band is absolutely empty, and the valence band completely filled (i.e. $n=p=0$).

Electrons from the donor level will try to go over to the lower acceptor level. If $N_d=N_a$ an equal number of donor and acceptor ions will be found in the semiconductor ($N_d^+=N_a^-$) and all donor states will be empty while all acceptor states will be occupied. As temperature rises the conduction band will receive electrons from the valence band and from the acceptor level E_a (there being no electrons on the donor level). Since, however, the energy gap

$E_c - E_a$ is almost equal to the forbidden band the increase in the concentration n will be about the same as in an intrinsic semiconductor. The Fermi level will be close to the middle of the forbidden band as in an intrinsic semiconductor:

$$F = \frac{E_d + E_a}{2} \quad (T = 0). \tag{33.1}$$

Such a semiconductor is termed compensated since the impurities completely compensate each other and are unable to act as suppliers of electrons and holes. Using the electric neutrality equation we may write $n = p$ since $N_d = N_a$, and $n_d = p_a$. Such a semiconductor is intrinsic from the point of view of carrier concentration. In other respects, however, it behaves differently from a pure semiconductor, and in the first instance the difference is in the carrier mobility. This is quite understandable since *lattice imperfections are more numerous in a compensated semiconductor*. If impurity concentrations are unequal the compensation will not be complete. Let $N_d > N_a$. Now the quantity $N'_d = N_d - N_a$ will play the part of the donor impurity since $N_d - N'_d$ will be absorbed in compensating the acceptor impurity.

There is, however, some difference in behaviour between partly compensated and non-compensated semiconductors. To investigate this difference let us take another look at the electric neutrality equation:

$$n + n_d - p - p_a = N_d - N_a = N'_d. \tag{33.2}$$

At $T \to 0$ both n and p turn zero, and the electric neutrality equation assumes the form

$$n_d - p_a = N'_d \tag{33.3}$$

or

$$\frac{N_d}{\frac{1}{2} e^{\frac{E_d - F}{kT}} + 1} - \frac{N_a}{\frac{1}{2} e^{\frac{F - E_a}{kT}} + 1} = N_d - N_a = N'_d. \tag{33.4}$$

From physical considerations it may be assumed that $F > E_a$, and this results in $p_a = 0$ at $T = 0$. Hence

$$n_d = N_d - N_a = N'_d. \tag{33.5}$$

Now it will be easy to find the Fermi level position for this limiting case:

$$\frac{N_d}{\frac{1}{2} e^{\frac{E_d - F}{kT}} + 1} = N'_d \tag{33.6}$$

or

$$e^{\frac{E_d-F}{kT}} = 2\frac{N_d-N_d'}{N_d'} = 2\frac{N_a}{N_d-N_a}, \tag{33.7}$$

whence

$$F = E_d + kT \ln \frac{N_d-N_a}{2N_a}. \tag{33.8}$$

At $T=0$ $F=E_d$, i.e. *the Fermi level coincides with the donor level.* We may write for the electron concentration

$$\left.\begin{aligned} n &= N_c \cdot \frac{N_d-N_a}{2N_a} e^{-\frac{\Delta E_d}{kT}}, \\ n &= \frac{(N_d-N_a)N_c}{2N_a} e^{-\frac{\delta E_a}{kT}} \end{aligned}\right\} \tag{33.9}$$

The activation energy δE_a *is equal to the ionization energy of the donor impurity* ΔE_d, *while in a semiconductor doped only with donors the activation energy is* $\frac{\Delta E_d}{2}$. The formulae (33.8) and (33.9) are valid in the lowest temperature range including $T=0$. They are, however, applicable at $T \neq 0$, as well. N_a in expressions (33.8-9) should not be zero or equal to N_d: $0 < N_a < N_d$.

For $N_a > N_d > 0$ we may write an expression completely analogous to (33.9):

$$F = E_a - kT \ln \frac{N_a-N_d}{2N_d} \tag{33.10}$$

and

$$p = \frac{(N_a-N_d)N_v}{2N_d} e^{-\frac{\Delta E_a}{kT}}; \quad \delta E_a = \Delta E_a. \tag{33.11}$$

The expression (33.8) for F shows that as temperature rises *the Fermi level moves upwards from* $F=E_d$ *at* $T=0$, *or downwards depending upon the relation between* N_a *and* N_d. For $N_d = 3N_a$, $F=E_d$ and F is independent of temperature (naturally, only insofar as the initial equations remain valid). For $N_d > 3N_a$ the Fermi level rises, the rate being the greater the less is N_a. For $N_d < 3N_a$ the Fermi level sinks as the temperature rises as is shown in Fig. 55.

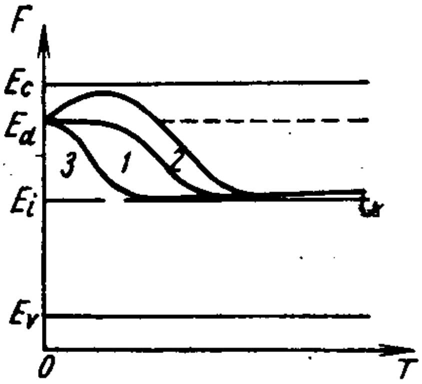


Fig. 55. The temperature dependence of the Fermi level in a semiconductor doped with both the donor and the acceptor impurities:

$1 - N_d = 3N_a$; $2 - N_d > 3N_a$; $3 - N_d < 3N_a$

If the expression for F in a partly compensated semiconductor is available it will be possible to obtain a more general expression valid in a greater temperature range by using the following electric neutrality equation:

$$n+n_d=N_d-N_a=N'_d; \quad n=N'_d-n_d=N'^+_d. \tag{33.12}$$

Taking into account the solution of this equation for the low temperature range we may write

$$F=\frac{E_c+E_d}{2}+\frac{kT}{2}\ln\frac{N_d}{2N_c}=\frac{E_c+E_d}{2}+\frac{kT}{2}\ln\frac{N_d-N_a}{2N_c} \tag{33.13}$$

and

$$n=\sqrt{\frac{(N_d-N_a)N_c}{2}}\,e^{-\frac{\Delta E_d}{2kT}}. \tag{33.14}$$

Contrary to equations (33.8) and (33.9) which are valid for $T=0$, as well, the equations (33.13) and (33.14) are of no use for $T=0$ since n is comparable with p_a, $N_c\cong 0$ while $N_a\neq 0$. The lower limit of validity of this formula may be defined by the relation

$$n=p_a; \quad (p\ll p_a) \quad \text{for} \quad F\cong E_d. \tag{33.15}$$

The value of T at which (33.8) should be replaced by (33.13) is easily obtained from here:

$$kT=\frac{\Delta E_0-\Delta E_d}{\ln N_a-\ln 2N_c}\cong\frac{\Delta E_0}{\ln N_a}. \tag{33.16}$$

It follows that the greater is N_a the greater is the temperature range 0-T where the applicability of the formulae (33.15) is limited and where the expression (33.8) holds.

Thus, *in the presence of a compensating impurity the rate of change of the Fermi level position with temperature increases.*

Consider now the general case of the dependence of the Fermi level on temperatute in the low temperature range for different concentration ratios of N_d and N_a. Of particular interest will be the case of $N_a\cong N_d$.

The electric neutrality equation

$$n+n_d-p-p_a=N_d-N_a \tag{33.17}$$

may be simplified for low temperatures by omitting n and p. This is the more justifiable the less are T and the difference N_d-N_a. The case of $N_d\gg N_a$ has been discussed above. The equation (33.17) in this case may be written in the form:

$$N_d-n_d=N_d^+=p_d=N_a-p_a=N_a^-=n_a, \tag{33.18}$$

or

$$\frac{N_d}{2e^{-\frac{F-E_d}{kT}}+1}=\frac{N_a}{2e^{\frac{E_a-F}{kT}}+1}. \tag{33.19}$$

Introduce the notation

$$e^{\frac{F}{kT}}=x;\quad \frac{1}{2}e^{\frac{E_d}{kT}}=D;\quad 2e^{\frac{E_a}{kT}}=A;\quad \frac{N_d}{N_a}=r;\quad r-1=s \tag{33.20}$$

and write the equation (33.19) using the new notation:

$$r\left(\frac{A}{x}+1\right)=\left(\frac{x}{D}+1\right) \tag{33.21}$$

or

$$x^2-sDx-rAD=0, \tag{33.22}$$

whence

$$x=\frac{sD}{2}\left(1\pm\sqrt{1+\frac{4rA}{s^2D}}\right)=\frac{D}{2}\left(\ \pm\sqrt{s^2+4r\frac{A}{D}}\right). \tag{33.23}$$

Since F is real, $x>0$, and both solutions should be used depending on the sign of s.

For $s>0$, i.e.

$$r=\frac{N_d}{N_a}>1,\quad \text{or}\quad N_d>N_a,$$

$$x=\frac{sD}{2}\left(1+\sqrt{1+\frac{4rA}{s^2D}}\right). \tag{33.24}$$

For $s<0$, i.e.

$$r=\frac{N_d}{N_a}<1,\quad \text{or}\quad N_d<N_a,$$

$$x=\frac{sD}{2}\left(1-\sqrt{1+\frac{4rA}{s^2D}}\right). \tag{33.25}$$

For $s=0$, i.e.

$$r=1,\quad \text{or}\quad N_d=N_a,$$

$$x=\sqrt{AD}=\left(2e^{\frac{E_a}{kT}}\cdot\frac{1}{2}e^{\frac{E_d}{kT}}\right)^{1/2}=e^{\frac{E_a+E_d}{2kT}}=e^{\frac{F}{kT}}. \tag{33.26}$$

In the latter case the expression for F is

$$F=\frac{E_d+E_a}{2}. \tag{33.27}$$

We have obtained it before from obvious considerations.

For positive s it follows from (33.24):

$$F=kT\ln\left\{\frac{sD}{2}\left(1+\sqrt{1+\frac{4rA}{s^2D}}\right)\right\}. \tag{33.28}$$

Taking into account that $\frac{A}{D}=4e^{-\frac{E_d-E_a}{kT}}\rightarrow 0$ at $T\rightarrow 0$ we obtain for large s and $\frac{4rA}{s^2D}\ll 1$:

$$F=kT\ln\frac{sD}{2}\cdot 2=kT\ln\left\{\frac{N_d-N_a}{N_a}\frac{1}{2}e^{\frac{E_d}{kT}}\right\}=E_d+kT\ln\frac{N_d-N_a}{2N_a}; \tag{33.29}$$

for small s and $\frac{4rA}{s^2D}\gg 1$

$$F=\frac{E_a+E_d}{2}. \tag{33.30}$$

If $s<0$ it follows from (33.25):

$$F=kT\ln\left\{\frac{|s|D}{2}\left[\sqrt{1+\frac{4rA}{s^2D}}-1\right]\right\}; \tag{33.31}$$

for large $|s|$ and $\frac{4rA}{s^2D}\ll 1$

$$\begin{aligned}F&=kT\ln\left\{\frac{|s|D}{2}\left[1+\frac{2rA}{s^2D}+\ldots-1\right]\right\}\cong\\ &\cong kT\ln\frac{rA}{|s|}=E_a+kT\ln\frac{N_a-N_d}{2N_d},\end{aligned} \tag{33.32}$$

which is in full accord with (33.29) for $s>0$; for small $|s|$ and $\frac{4rA}{s^2D}\gg 1$

$$F=kT\ln\left\{\frac{|s|D}{2}\sqrt{\frac{4rA}{s^2D}}\right\}=kT\ln\sqrt{AD}=\frac{E_a+E_d}{2} \tag{33.33}$$

in full accord with (33.30).

The Fermi level regarded as a function of s at a constant temperature is fully described by the expressions (33.28) and (33.31).

From (33.29) and (33.32) we obtain for the electron and hole concentrations:

$$n=\frac{(N_d-N_a)N_c}{2N_a}e^{-\frac{\Delta E_d}{kT}}, \tag{33.34}$$

$$p=\frac{(N_a-N_d)N_v}{2N_d}e^{-\frac{\Delta E_a}{kT}}. \tag{33.35}$$

The slope of the $\ln n$ vs $\frac{1}{T}$ *graph is determined by the value of* $\frac{\Delta E_d}{k}$ *and* $\frac{\Delta E_a}{k}$ *and is twice as large as the slope in the case of a semiconductor doped with impurity of one kind.*

Summary of Sec. 33

1. By changing the concentrations of the donor N_d and acceptor N_a impurities it is possible to vary the concentrations of free electrons n and holes p within a wide range. The product of the concentrations n and p is, however, independent of the concentration and sort of impurities as long as the semiconductor remains non-degenerate. This product is equal to

$$np = n_i^2. \tag{33.1s}$$

2. The semiconductor in which $N_d = N_a$ is termed compensated. The Fermi level in it lies close to the middle of the forbidden band.
3. If there is some difference between the impurity concentrations the Fermi level at $T = 0$ will coincide with the level of the more abundant impurity. The temperature of transition to intrinsic conductivity is the lower the closer are the impurity concentrations.
4. If the concentrations N_d and N_a differ greatly the semiconductor will behave as one doped with impurity of one kind.

34. DEGENERATE SEMICONDUCTOR

In Sec. 29 we defined a degenerate semiconductor as one in which the Fermi level lies inside an energy band no less than 5 kT from its boundary. Free charge carrier concentration in this case *is independent of temperature*. Now we are about to demonstrate that *degeneracy may be attained only through heavy doping*.

As was stated above, in an extrinsic semiconductor the Fermi level, as temperature increases, approaches the corresponding energy band.

Find how the position of the maximum of F changes with temperature. We obtain from (32.14)

$$\frac{dF}{dT} = \frac{k}{2} \ln \frac{N_d}{2N_c} - \frac{kT}{2} \cdot \frac{N_d}{2N_c} \cdot \frac{2}{N_d} \cdot \frac{dN_c}{dT} = 0, \tag{34.1}$$

or

$$\ln \frac{N_d}{2N_c} - \frac{T}{N_c} \cdot \frac{dN_c}{dT} = 0. \tag{34.2}$$

But since

$$\frac{dN_c}{dT} = \frac{3}{2} \frac{N_c}{T}, \tag{34.3}$$

the condition for the extremum (34.1) or (34.2) assumes the form

$$\ln \frac{N_d}{2N_c} = \frac{3}{2}; \quad \frac{N_d}{2N_c} = e^{3/2}, \tag{34.4}$$

i.e. the temperature T_{max} at which F attains its maximum value $F = F_{max}$ is determined by the condition

$$N_c(T_{max}) = \frac{N_d}{2e^{3/2}}. \qquad (34.5)$$

Taking account of expression (29.22) for N_c we obtain the following expression for T_{max}:

$$T_{max} = 8.15\left(\frac{m}{m_d^*}\right)\cdot\left(\frac{N_d}{10^{18}}\right)^{2/3}. \qquad (34.6)$$

The expression (34.6) shows that the temperature T_{max} *increases with the increase of impurity concentration as* $N_d^{2/3}$. Find the value of F_{max}:

$$F_{max} = \frac{E_c + E_d}{2} + \frac{kT_{max}}{2}\ln\frac{N_d}{2N_c(T_{max})} = \frac{E_c + E_d}{2} + \frac{3}{4}kT_{max}, \qquad (34.7)$$

or

$$F_{max} = \frac{E_c + E_d}{2} + 5.3\times10^{-4}\left(\frac{m}{m_d^*}\right)\cdot\left(\frac{N_d}{10^{18}}\right)^{2/3}. \qquad (34.8)$$

In (34.8) and below k is expressed in electron-volts. *The concentration $N_d^{(cr)}$ for which F coincides with the bottom of the conduction band is termed critical.* It may be found from the condition

$$F_{max} = E_c = \frac{E_c + E_d}{2} + 5.3\times10^{-4}\left(\frac{m}{m_d^*}\right)\cdot\left(\frac{N_d^{(cr)}}{10^{18}}\right)^{2/3}, \qquad (34.9)$$

or

$$N_d^{(cr)}(\text{cm}^{-3}) = 10^{22.5}\left(\frac{m_d^*}{m}\right)^{3/2}[\Delta E_d(\text{eV})]^{3/2}. \qquad (34.10)$$

Assess the critical concentration $N_d^{(cr)}$ for $m_d^* = m$ and $\Delta E_d = 0.03$ eV. Substituting the values m_d^* and ΔE_d into (34.10) we obtain $N_d^{(cr)} = 1.6\times10^{20}$ cm^{-3}. If m_d^*/m is put at 0.3, $N_d^{(cr)}$ will be 2.5×10^{19} cm^{-3}.

Thus, *the critical concentration is very sensitive to the impurity ionization energy and to the effective mass.* In A$^{\text{III}}$B$^{\text{V}}$ compounds the critical concentration may be much lower than 10^{19} cm^{-3}. Indeed, if we put $m_d^*/m = 10^{-2}$ and $\Delta E_d = 0.0001$ eV we obtain for $N_d^{(cr)}$ the value $N_d^{(cr)} = 10^{13}$ cm^{-3} observed in indium antimonide. With the help of the critical concentration it is possible to assess the impurity concentration at which degeneracy of the semiconductor sets in, since when $F = E_c$ the semiconductor ceases to be non-degenerate. However, it is not yet degenerate in the sense that charge carrier concentration within some temperature range becomes independent of temperature. Greater impurity concentration is required for this. The position of the Fermi level and the charge carrier concentration may, in principle, be assessed from the electric neutrality equation from which for $N_a = 0$ p and p_a may be omitted. We will have $n + n_d = N_d$, or $n = p_d = N_d^+$. Substituting the expressions for n and p_d we write:

$$\frac{2N_c}{\sqrt{\pi}}\Phi_{1/2}(\xi) = \frac{N_d}{2e^{\frac{F-E_d}{kT}}+1}. \tag{34.11}$$

For a completely degenerate semiconductor we may in the expression for p_d neglect unity as compared with the exponent. Substituting the expression for $\Phi_{1/2}(\xi)$ we write:

$$\frac{4N_c}{3\sqrt{\pi}}\left(\frac{F-E_c}{kT}\right)^{3/2} = \frac{N_d}{2}e^{-\frac{F-E_d}{kT}} \tag{34.12}$$

We may write the following equation from (34.12):

$$\xi^{3/2}e^{\xi} = \left(\frac{3\sqrt{\pi}}{8}\cdot\frac{e^{-\frac{\Delta E_d}{kT}}}{N_c}\right)N_d. \tag{34.13}$$

n may be found after ξ has been determined from (34.13) as a function of N_d.

It must, however, be noted that there is little sense in the expressions (34.12) and (34.13) since *at such great concentrations as are needed for degeneracy the impurity level turns into a band which merges with the conduction band.* The impurity band is only partially occupied. The result is that degeneracy does not vanish even at very low temperatures since impurity band conduction mechanism remains active. *Because of the merger of the bands degeneracy persists in a wide temperature range*, this range extending for some intermetallic compounds from room to liquid hydrogen temperatures. It must be noted besides, that *owing to impurity band formation impurity ionization energy decreases with the increase in impurity concentration, and this in turn, leads to the decrease in the concentration.*

Sometimes the need may arise to calculate the position of the Fermi level of a completely degenerate semiconductor from a known charge carrier concentration:

$$F - E_c = \left(\frac{h^2}{2m_{nd}^*}\right)\left(\frac{3}{8\pi}\right)^{2/3}\cdot n^{2/3}, \tag{34.14}$$

$$E_v - F = \left(\frac{h^2}{2m_{pq}^*}\right)\left(\frac{3}{8\pi}\right)^{2/3}\cdot p^{2/3}. \tag{34.15}$$

Electron or hole concentration for this purpose may be obtained from experiment.

Degenerate semiconductors are used in the manufacture of such devices as tunnel diodes and semiconductor lasers. Besides, they are important for theoretical studies. Heavily doped semiconductors present principal difficulties for theoretical analysis because high electron concentration leads to a strong interaction between the electrons and the donor ion, as a result of which the latter is screened. This screening of the donor field decreases the ionization energy almost to zero, and this makes the talk about the broadening of impurity levels irrelevant. To study heavily doped semiconductors such methods as the method of Green's functions should be used.

Note in conclusion that *in a heavily doped semiconductor the density of states in a band is greatly distorted by the impurity. This distortion remains even when impurity concentration decreases. It leads to the blurring of sharp band boundaries, to the appearance of a density of states "tail" which attenuates exponentially and reaches as far as the neighbouring band.*

Summary of Sec. 34

1. The increase in impurity concentration leads to an increase in the charge carrier concentration, and this, in turn, to the decrease in the distance between the energy band and the Fermi level.

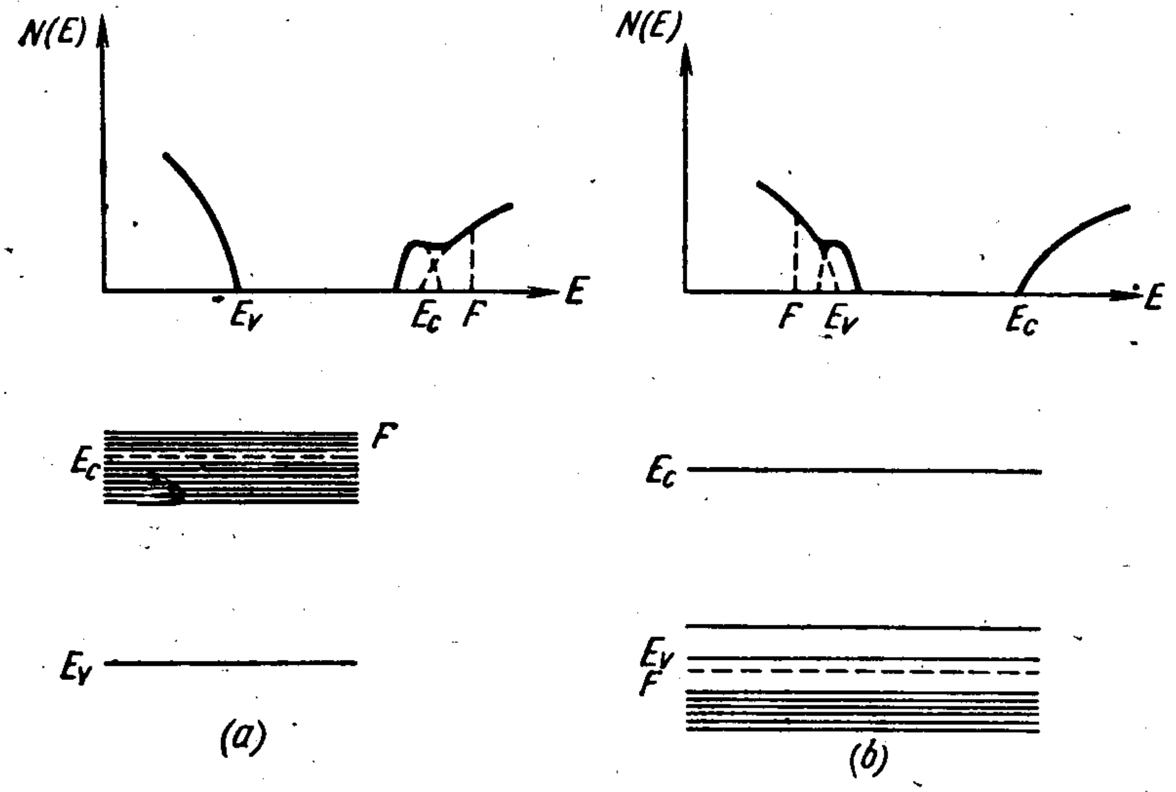


Fig. 56. The density of states and the position of the Fermi level in a degenerate electron (*a*) and hole-type (*b*) semiconductors

For a definite impurity concentration $N^{(cr)}$ the Fermi level coincides with the energy extremum in the band. The concentration $N^{(cr)}$ is termed critical. When impurity concentration exceeds $N^{(cr)}$ the semiconductor becomes partly or completely degenerate.

2. The term lower degeneracy temperature T^{l}_{deg} applies to the temperature at which, as it is raised from $T=0$, the Fermi level enters an energy band. At $T<T^{l}_{deg}$ the charge carriers are "frozen out" from the energy band to the impurity level.

3. The term upper degeneracy temperature T^{up}_{deg} applies to the temperature at which the electron gas becomes classical. It is defined by the condition

$$kT^{up}_{deg}=F-E_c; \quad kT^{up}_{deg}=E_v-F. \tag{34.1s}$$

The charge carrier concentration in this case may, however, remain constant if the temperature is not high enough for transition to intrinsic conductivity to take place.

4. There may be no lower degeneracy temperature in a semiconductor if the impurity level broadens into an impurity band as impurity concentration increases. The impurity band my be only partially filled, and in this case the charge carriers in it will be free at any temperature, including $T=0$, even if there is an energy slit between the energy and the impurity bands. At sufficiently high impurity concentrations the bands may overlap. In any case doping of a semiconductor results in a broadening of sharp energy band boundaries (Fig. 56).

35. DENSITY OF STATES IN A MAGNETIC FIELD

As was shown in Sec. 23, in a strong magnetic field the electron energy spectrum experiences a sharp change due to the appearance of Landau levels:

$$E=\frac{\hbar^2k_z^2}{2m^*}+\hbar\omega_0\left(n+\frac{1}{2}\right); \quad E_c=0, \tag{35.1}$$

where $n=0, 1, 2, \ldots$ is the quantum number, $\omega_0=\frac{eB}{m^*}$, the cyclotron frequency. In this case the concept of constant-energy surfaces in the Brillouin zone becomes meaningless since E does not depend on k_x and k_y. It must, however, be kept in mind that for $B\to 0$ the expression (35.1) must coincide with the expression

$$E=\frac{\hbar^2k_x^2}{2m^*}+\frac{\hbar^2k_y^2}{2m^*}+\frac{\hbar^2k_z^2}{2m^*}. \tag{35.2}$$

In other words, symbolically

$$\lim_{B\to 0}\hbar\omega_0(n+1/2)=\frac{\hbar^2(k_x^2+k_y^2)}{2m^*}. \tag{35.3}$$

For a semiconductor in a magnetic field the electron state may be defined with the aid of the dynamical quantities k_z, n, ω_0 and s_z.

Find the expression for the density of states in terms of energy $N_B(E)$. The interval E, $E+dE$ contains

$$2dS = 2N_B(E)\,dE \tag{35.4}$$

states. Obviously, the total number of states must be $2Ng$, where N is the number of atoms in the crystal, g—band degeneracy factor, 2—a coefficient due to spin degeneracy:

$$2\int_{E_{min}}^{E_{max}} N_B(E)\,dE = 2N\cdot g. \tag{35.5}$$

For fixed values of n and ω_0 the number of states in an interval dk_z (for different k_z) per unit crystal volume is

$$dS_{kz} = \frac{dk_z}{2\pi}. \tag{35.6}$$

Using (35.1) find the relation between dk_z and dE:

$$k_z = \left(\frac{2m^*}{\hbar^2}\right)^{1/2}[E - \hbar\omega_0(n+1/2)]^{1/2}, \tag{35.7}$$

$$dk_z = \frac{1}{2}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}[E - \hbar\omega_0(n+1/2)]^{-1/2}\cdot dE. \tag{35.8}$$

In compliance with (35.6) one may write:

$$dS_{k_z} = \frac{dk_z}{2\pi} = \frac{1}{4\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}[E - \hbar\omega_0(n+1/2)]^{-1/2}dE. \tag{35.9}$$

If it is now desired to write the expression for the density of states in energy, the following consideration should be taken into account. There are several intervals dk_z for a fixed energy interval E, $E+dE$, this being evident from the expression (35.8) and Fig. 57*a*.

Indeed, there are ν parabolae which correspond to the energy E, where ν is determined by the relation

$$\nu \leqslant \frac{E - \frac{\hbar\omega_0}{2}}{\hbar\omega_0} < \nu + 1. \tag{35.10}$$

n may change from zero to ν, therefore, there are ν intervals dk_z to correspond to the energy interval E, $E+dE$, and for this reason to find dS_{k_z} one should sum up all the states of ν intervals of dk_z:

$$dS_{k_z}=\left\{\frac{1}{4\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}\sum_{n=0}^{\nu}[E-\hbar\omega_0(n+1/2)]^{-1/2}\right\}dE. \quad (35.11)$$

Since the expression for energy contains the term $\hbar\omega_0=\frac{e\hbar}{m^*}B$, the density of states should moreover be dependent on B. Suppose it is determined by some function $g(B)$ the form of which we intend to find from its behaviour in the limiting case $B \to 0$.

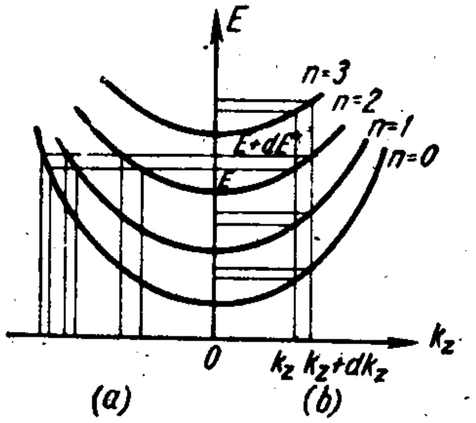


Fig. 57. The relation between the energy interval dE and the interval of dk_z values

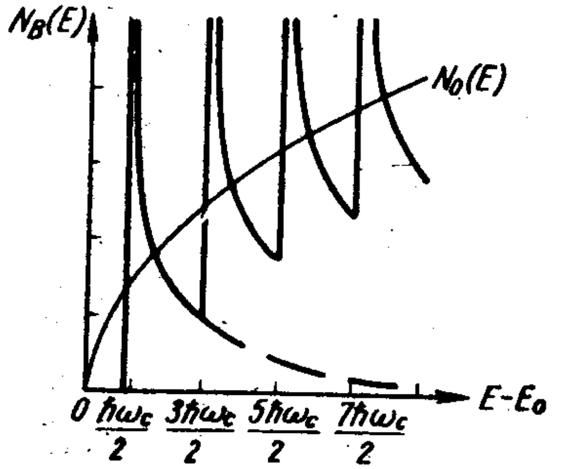


Fig. 58. The density of states in a magnetic field

Comparing (35.11) with (35.4) and taking account of the function $g(B)$ we may write the expression for the density of states:

$$N_B(E)=\frac{g(B)}{4\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}\sum_{n=0}^{\nu}[E-\hbar\omega_0(n+1/2)]^{1/2}=\sum_{n=0}^{\nu}N_B^{(n)}(E). \quad (35.12)$$

Consider one of the addends of (35.12)

$$N_B^{(n)}(E)=\frac{g(B)}{4\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}[E-\hbar\omega_0(n+1/2)]^{-1/2}. \quad (35.13)$$

Since n is fixed, the density $N_B^{(n)}(E)$ for $E \gg \hbar\omega_0(n+1/2)$ will be small and will behave as $\frac{1}{\sqrt{E}}$.

Since the value of $N_B^{(n)}(E)$ must be real, $N_B^{(n)}(E)$ is determined only for the energy $E > \hbar\omega_0(n+1/2)$. For $E \to \hbar\omega_0(n+1/2)$, $N_B^{(n)}(E) \to \infty$. The graph of one of the $N_B^{(n)}(E)$ dependences corresponding to $n=0$ is shown in Fig. 58 by a solid and a dashed line.

The total density $N_B^{(n)}(E)$ is a sum of ν identical hyperbolae displaced along the energy axis by an integral number of $\hbar\omega_0$. At points $E=\hbar\omega_0(n+1/2)$ the density of states $N_B(E)$ becomes infinite. The meaning of this is quite obvious. When discrete levels are formed from a continuous spectrum, the density of states must be of a δ-nature so that the total number of states remains finite. But a singularity of the form (35.13) for any finite energy interval does, in fact, contain a finite number of states:

$$\int\limits_{\hbar\omega_0(n+1/2)}^{E_0} N_B^{(n)}(E)\,dE = \frac{g(B)}{4\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2} \times$$

$$\times \int\limits_{\hbar\omega_0(n+1/2)}^{E_0} [E-\hbar\omega_0(n+1/2)]^{-1/2}\,dE =$$

$$= \frac{g(B)}{2\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}[E_0-\hbar\omega_0(n+1/2)]^{1/2}. \quad (35.14)$$

Hence, in any finite energy interval $\{\hbar\omega_0(n+1/2);\ \hbar\omega_0(n+1/2)+ +\Delta E\}$ there is a finite number of states

$$\frac{g(B)}{2\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}(\Delta E)^{1/2}. \quad (35.15)$$

An infinitely large density of states corresponds to the value of $k_z=0$; without the magnetic field at the point $k_z=0$ (and $k_x= =k_y=0$) $N(E)=0$. (To be definite we presume that there is an energy extremum at the point $\mathbf{k}=0$.)

The shape of the $N_B(E)$ graph in the vicinity of the nth singular point is determined only by the term $N_B^{(n)}(E)$, while away from these points the sum of ν hyperbolae must be taken into account. Evidently, the higher is E the greater is ν, and the greater is $N_B(E)$ in the interval of E between the singular points. This is clearly shown in Fig. 58. Here *the value of $N_B(E)$ to the left of every singular point is the greater the greater is* ν (*or* E). For $B\longrightarrow 0$ ν for any finite value of E will tend to infinity. However, this tendency is such that $\lim\limits_{\nu\to\infty}\nu\hbar\omega_0=\lim\limits_{B\to\infty}\nu\hbar\omega_0=E$ for arbitrary energy values. For $B\longrightarrow 0$

$$N_B(E)\longrightarrow N_0(E)=2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}E^{1/2}=N(E). \quad (35.16)$$

Now we can find the extreme value of $N_0(E)$ replacing the sum over n by an integral:

$$\sum_{n=0}^{\nu}[E-\hbar\omega_0(n+1/2)]^{-1/2}\longrightarrow\int\limits_0^{\nu}[E-\hbar\omega_0(n+1/2)]^{-1/2}\,dn=$$

$$=-\frac{1}{\hbar\omega_0}\int\limits_{E-\frac{\hbar\omega_0}{2}}^{0}x^{-1/2}\,dx=\frac{2}{\hbar\omega_0}\left(E-\frac{\hbar\omega_0}{2}\right)^{1/2}. \quad (35.17)$$

For weak fields $\sum\limits_{n=0}^{\nu}\left[E-\hbar\omega_0\left(n+\frac{1}{2}\right)\right]^{-1/2}$ is transformed into $\frac{2}{\hbar\omega_0}\left(E-\frac{\hbar\omega_0}{2}\right)^{1/2}$

For $B\rightarrow 0$ $\omega_0\rightarrow 0$ and $\left(E-\frac{\hbar\omega_0}{2}\right)^{1/2}\rightarrow E^{1/2}$, but $\frac{1}{\hbar\omega_0}\rightarrow\infty$.

Consider now the function $g(B)$. If we put

$$g(B)=G\cdot\hbar\omega_0 \tag{35.18}$$

the sum (35.17) will become convergent. Find G.

$$N_B(E)=\frac{G\cdot\hbar\omega_0}{4\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}\sum_{n=0}^{\nu}[E-\hbar\omega_0(n+1/2)]^{-1/2}\rightarrow$$
$$\rightarrow\frac{G\hbar\omega_0}{4\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}\cdot\frac{2}{\hbar\omega_0}\left(E-\frac{\hbar\omega_0}{2}\right)^{1/2}\rightarrow N_0(E)=2\pi\times$$
$$\times\left(\frac{2m^*}{h^2}\right)^{3/2}E^{1/2}. \tag{35.19}$$

Comparing the limiting expressions for $N_B(E)$ and $N(E)$ we may write

$$\frac{G}{2\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}=2\pi\left(\frac{2m^*}{h^2}\right)^{3/2} \tag{35.20}$$

It follows from (35.20) and (35.18) that

$$G=\frac{4\pi m^*}{h^2}, \tag{35.21}$$

$$g(B)=G\cdot\hbar\omega_0=\frac{2m^*}{h}\omega_0=\frac{2eB}{h}, \tag{35.22}$$

and

$$N_B(E)=\frac{\hbar\omega_0}{2}\left[2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}\right]\sum_{n=0}^{\nu}[E-\hbar\omega_0(n+1/2)]^{-1/2}. \tag{35.23}$$

The expression (35.23) *represents the density of states for a crystal in a magnetic field.* In deriving it we have neglected the magnetic moment of the electron $\frac{e}{m^*}\mathrm{S}=\mu^*$ which accounts for additional energy of the electron

$$-(\mu^*\mathbf{B})=-\left(\frac{e\mathbf{B}}{m^*}\mathrm{S}\right)=-(\mathrm{S}\omega_0)=\pm\mu_0^*B=\pm\frac{\hbar\omega_0}{2}, \tag{35.24}$$

where

$$\mu_0^*=\frac{e\hbar}{2m^*}=\frac{m}{m^*}\mu_0 \tag{35.25}$$

is the Bohr magneton for an electron with the effective mass m^*; $\mu_0 = \frac{e\hbar}{2m}$ is the Bohr magneton of a free electron; and **S** is the spin vector the projection of which on **B** is $\pm\frac{\hbar}{2}$. Neglecting the energy level splitting we multiply each energy state by two.

Denoting the Fermi energy by $F(B)$ we write the Fermi-Dirac function

$$f = \frac{1}{e^{\frac{E-F(B)}{kT}}+1}. \tag{35.26}$$

For the electron concentration we obtain

$$n_n = 2\int_0^\infty N_B(E)\,f(E,\ T)\,dE = \hbar\omega_0\left[2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}\right]\times$$

$$\times\int_0^\infty\sum_{n=0}^{\nu}[E-\hbar\omega_0(n+1/2)]^{-1/2}\left[e^{\frac{E-F(B)}{kT}}+1\right]^{-1}dE. \tag{35.27}$$

Calculate this expression for two limiting cases—for degenerate and non-degenerate semiconductors.

In case of the *degenerate semiconductor* we replace the distribution function by a rectangular step, and, correspondingly, the expression for n_n assumes the form

$$n_n = \hbar\omega_0\left[2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}\right]\int_0^{F(B)}\sum_{n=0}^{\nu}[E-\hbar\omega_0(n+1/2)]^{-1/2}dE =$$

$$=\hbar\omega_0\left[2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}\right]\sum_{n=0}^{\nu}\int_{\hbar\omega_0(n+1/2)}^{F(B)}[E-\hbar\omega_0(n+1/2)]^{-1/2}dE=$$

$$=2\hbar\omega_0\left[2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}\right]\sum_{n=0}^{\nu}[F(B)-\hbar\omega_0(n+1/2)]^{1/2}. \tag{35.28}$$

In a weak magnetic field the sum can be replaced by an integral:

$$\hbar\omega_0\sum_{n=0}^{\nu}[F(B)-\hbar\omega_0(n+1/2)]^{1/2} \cong$$

$$\cong\hbar\omega_0\int_0^{\nu}[F(B)-\hbar\omega_0(n+1/2)]^{1/2}dn \cong \frac{2}{3}F^{3/2}(B), \tag{35.29}$$

$$n_n = \frac{8\pi}{3}\left(\frac{2m^*}{h^2}\right)^{3/2}F^{3/2}(B). \tag{35.30}$$

Using (35.30) we write the expression for $F(B)$:

$$F(B)=\frac{h^2}{2m^*}\left(\frac{3n_n}{8\pi}\right)^{2/3}=F_0. \tag{35.31}$$

In a strong magnetic field ν will be small. Consider a field in which only one level $n=0$ remains unexcited (*the quantum limit*). In this case only one term remains from the sum (35.28):

$$n_n=2\hbar\omega_0\left[2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}\right]\left[F(B)-\frac{\hbar\omega_0}{2}\right]^{1/2} \tag{35.32}$$

whence

$$F(B)=\frac{\hbar\omega_0}{2}+\frac{n_n^2}{16\pi^2\left(\frac{2m^*}{h^2}\right)^3(\hbar\omega_0)^2}=\frac{\hbar\omega_0}{2}\left[1+\frac{1}{18}\left(\frac{F_0}{\frac{\hbar\omega_0}{2}}\right)^3\right], \tag{35.33}$$

where F_0 is determined by the expression (35.31).

Consider a *non-degenerate semiconductor*. The electron concentration in it is

$$n_n=\hbar\omega_0\left[2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}\right]\times$$
$$\times e^{\frac{F(B)}{kT}}\sum_{n=0}^{\nu}\int\limits_{\hbar\omega_0(n+1/2)}^{\infty}[E-\hbar\omega_0(n+1/2)]^{-1/2}e^{-\frac{E}{kT}}dE. \tag{35.34}$$

The singularity of the function under the integral sign at the point $E=\hbar\omega_0(n+1/2)$ is of the δ-type and does not cause the integral to diverge, this being easily checked by direct calculation. Indeed, consider the integral

$$\int\limits_a^{\infty}(x-a)^{-1/2}e^{-\frac{x}{b}}dx=2(x-a)^{1/2}e^{-\frac{x}{b}}\Big|_a^{\infty}+$$
$$+\frac{2}{b}\int\limits_a^{\infty}(x-a)^{1/2}\mathrm{e}^{-\frac{x}{b}}dx=\frac{2}{b}b^{3/2}e^{-\frac{a}{b}}\int\limits_a^{\infty}\frac{(x-a)^{1/2}}{\sqrt{b}}e^{-\frac{(x-a)}{b}}d\left(\frac{x}{b}\right)=$$
$$=2b^{1/2}e^{-\frac{a}{b}}\int\limits_0^{\infty}y^{1/2}e^{-y}dy=2b^{1/2}e^{-\frac{a}{b}}\Gamma\left(\frac{3}{2}\right)=\sqrt{\pi b}\,e^{-\frac{a}{b}}. \tag{35.35}$$

We see from (35.35) that the result of the integration is the value of the function $e^{-\frac{x}{b}}$ at the point $x=a$, i. e. $N_B^{(n)}(E)$ actually behaves like a δ-function. Write the expression for n_n taking account of the fact that $b=kT$; $a=\hbar\omega_0(n+1/2)$:

$$n_n=\hbar\omega_0\left[2\pi\left(\frac{2m^*}{h^2}\right)^{3/2}\right]\sqrt{\pi kT}\,e^{\frac{F(B)}{kT}}\sum_{n=0}^{\infty}e^{-\frac{\hbar\omega_0(n+1/2)}{kT}} \tag{35.36}$$

But

$$\sum_{n=0}^{\infty} e^{-\frac{\hbar\omega_0}{kT}n} = \frac{1}{1-e^{-\frac{\hbar\omega_0}{kT}}}, \tag{35.37}$$

since $\left(e^{-\frac{\hbar\omega_0}{kT}}\right)^n$ forms a geometrical progression. If we take into account that

$$\frac{e^{-\frac{\hbar\omega_0}{2kT}}}{1-e^{-\frac{\hbar\omega_0}{kT}}} = \frac{1}{e^{\frac{\hbar\omega_0}{2kT}} - e^{-\frac{\hbar\omega_0}{2kT}}} = \left[2\sinh\left(\frac{\hbar\omega_0}{2kT}\right)\right]^{-1}, \tag{35.38}$$

we may finally write for n_n

$$n_n = \left(\frac{\hbar\omega_0}{2kT}\right) N_c \frac{e^{\frac{F(B)}{kT}}}{\sinh\left(\frac{\hbar\omega_0}{2kT}\right)}. \tag{35.39}$$

For a known concentration n_n the expression for the Fermi level $F(B)$ may be found:

$$F(B) = kT \ln\left\{\frac{n_n}{N_c}\frac{\sinh\left(\frac{\hbar\omega_0}{2kT}\right)}{\left(\frac{\hbar\omega_0}{2kT}\right)}\right\}. \tag{35.40}$$

In a weak magnetic field

$$\sinh\left(\frac{\hbar\omega_0}{2kT}\right) \cong \left(\frac{\hbar\omega_0}{2kT}\right), \tag{35.41}$$

$$F = kT \ln\frac{n_n}{N_c}. \tag{35.42}$$

Since F in (35.42) is measured from E_c (i. e. $E_c = 0$), (35.42) assumes the form

$$n_n = N_c e^{-\frac{E_c - F}{kT}}, \tag{35.43}$$

which exactly coincides with the expression (29.11s).

In a strong magnetic field

$$\sinh\left(\frac{\hbar\omega_0}{2kT}\right) \cong e^{\frac{\hbar\omega_0}{2kT}}, \tag{35.44}$$

$$n_n = \left(\frac{\hbar\omega_0}{2kT}\right) N_c e^{\frac{F(B)-\frac{\hbar\omega_0}{2}}{kT}} \tag{35.45}$$

or

$$n_n \doteq \left(\frac{\hbar\omega_0}{2kT}\right) N_c e^{-\frac{E_c+\frac{\hbar\omega_0}{2}-F(B)}{kT}}. \qquad (35.46)$$

The quantity

$$E_c + \frac{\hbar\omega_0}{2} = E_c(B) \qquad (35.47)$$

represents the lowest ($n=0$) energy level in an applied magnetic field. It may, therefore, be said that *magnetic field tends to remove the degeneracy and to increase the forbidden band width.*

Summary of Sec. 35

1. In magnetic field the density of states experiences substantial changes. It turns infinite in the "apexes" of the Landau parabolae and decreases with the increase in energy for each Landau parabola as $\frac{1}{\sqrt{E}}$. It is this property of the density of states in a magnetic field that justifies the concept of discrete Landau levels.

2. The expressions for the charge carrier concentration and for the Fermi level in a weak magnetic field coincide with corresponding expressions for a semiconductor in the absence of a magnetic field. In a strong magnetic field these expressions are notably different.

3. The charge carrier concentration considered in this chapter is due to thermal generation, i. e. the charge carriers are generated at the expense of the energy of thermal vibrations of the lattice with which the electron (or hole) gas is in thermodynamical equilibrium. For this reason such charge carriers are termed equilibrium carriers, and their concentration is termed equilibrium charge carrier concentration.

Chapter IV

KINETIC PHENOMENA IN SEMICONDUCTORS

36. BOLTZMANN'S KINETIC EQUATION

Kinetic, or transport, phenomena is the term applied to physical phenomena resulting from the motion of charge carriers under the action of internal or external fields or temperature gradients. They include electric and thermal conductivities, galvanomagnetic, thermomagnetic and thermoelectric phenomena. Kinetic phenomena are the basis of photoelectric and photomagnetic effects. The concept of particles moving under the action of external forces is sufficient for a qualitative description of kinetic phenomena. However, such particle motion models are inadequate for quantitative description. To obtain correct expressions for quantities which describe kinetic phenomena one should make use of methods of a more general nature which take account of the part played by charge carriers in different states. Specifically, the method of the Boltzmann kinetic equation is that sufficiently powerful method of theoretical investigation of kinetic effects which discribes the change in the states of particles under the influence of various factors.

In an ideal crystal the state ψ_k (r) remains unchanged for an infinitely long time. Because of this the distribution function of electrons among the states $f(\mathbf{r}, \mathbf{k})$, too, remains unchanged. When an external force field $V(\mathbf{r})$ is applied to the electron system of an ideal crystal, the state of each electron in the Brillouin zone changes according to the equation

$$\frac{d\mathbf{P}}{dt} = -\nabla V; \quad \frac{d\mathbf{k}}{dt} = \frac{1}{\hbar}\mathbf{F}_a, \qquad (36.1)$$

or

$$\mathbf{P}(t) = \int_0^t \mathbf{F}_a(\xi)\, d\xi + \mathbf{P}_0(0). \qquad (36.2)$$

The variation of the quasimomentum during the time t is independent of $\mathbf{P}_0$ and determited only by the momentum of the force

$$\Delta \mathbf{P} = \mathbf{P}(t) - \mathbf{P}_0(0) = \int_0^t \mathbf{F}_a(\xi)\, d\xi. \tag{36.3}$$

The relation (36.3) shows that if at some moment $t=0$ the electrons occupied states described by the function $f(\mathbf{r}, \mathbf{k}) = f\left(\mathbf{r}, \frac{\mathbf{P}_0}{\hbar}\right)$, at the moment t the distribution of electrons among the states must be determined by another function

$$f\left[\mathbf{r}; \ \mathbf{k}(t) - \frac{1}{\hbar}\int_0^t \mathbf{F}_a(\xi)\, d\xi\right] = \tilde{f}(\mathbf{r}, \mathbf{k}). \tag{36.4}$$

In other words, external (relative to the lattice periodic field) forces bring about the change in the electron distribution among states.

Write the equation which determines the change of the distribution function in time, i.e. write its full time derivative

$$\frac{df(\mathbf{r}, \mathbf{k}, t)}{dt} = \frac{\partial f}{dt} + \frac{\partial f}{\partial \mathbf{r}} \cdot \frac{d\mathbf{r}}{dt} + \frac{\partial f}{\partial \mathbf{k}} \cdot \frac{d\mathbf{k}}{dt} =$$

$$= \frac{\partial f}{\partial t} + (\nabla_r f \cdot \mathbf{v}) + \frac{1}{\hbar}(\nabla_k f \cdot \mathbf{F}_a). \tag{36.5}$$

We took into account (36.1) and the fact that $\frac{d\mathbf{r}}{dt} = \mathbf{v}$ is the eleetron velocity. We may write, in compliance with the Liouville theorem about the invariance of the phase volume for a system moving along the phase paths or on account of the conservation of the number of states, that

$$\frac{df}{dt} = 0, \tag{36.6}$$

or

$$-\frac{\partial f}{\partial t} = (\nabla_r f \cdot \mathbf{v}) + \left(\nabla_k f \cdot \frac{\mathbf{F}_a}{\hbar}\right). \tag{36.7}$$

The equation (36.7) shows that the *variation of the distribution function at each point of the phase space* ($\mathbf{r}$, $\mathbf{k}$) *with time is caused by the motion of particles in the normal space and in the space of the wave vector.*

The force $\mathbf{F}_a$ is determined both by external macroscopic fields and by all imperfections of the lattice field—by vacancies, impurity atoms or ions, and thermal lattice vibrations.

In most practically important cases it is desired to know how the solid behaves in external macroscopic fields. All forces due to any localized imperfection of the periodic field are in this case internal forces for the crystal in question and must be regarded as belonging to a special class. We will divide the forces $\mathbf{F}_a$ into two classes: one, denoted by $\mathbf{F}$, will comprise forces due to macroscopic external fields, and the other, denoted by $\mathbf{F}_d$, forces due to localized lattice field imperfections. The external $\mathbf{F}$ and internal $\mathbf{F}_d$ forces act in a quite different fashion. The external forces $\mathbf{F}$ cause the particles to move in the quasimomentum and the co-ordinate spaces. Indeed, for every particle

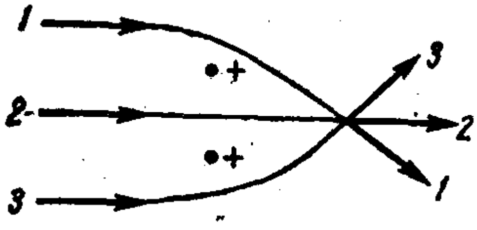


Fig. 59. The influence of the ionic fields on the variation of the quasimomentum of the electrons

$$\Delta \mathbf{P} = \int_0^t \mathbf{F}(\xi)\, d\xi = \mathbf{F}t, \qquad (36.8)$$

if $\mathbf{F}$ is independent of time. Exactly the same expression may be written for $\mathbf{F}_d$:

$$\Delta \mathbf{P} = \int_0^t \mathbf{F}_d(\xi)\, d\xi. \qquad (36.9)$$

However, at every point $\mathbf{F}_d$ is the resultant of a vast number of localized fields and therefore very sharply and intricately dependent on the co-ordinate of the electron or of the hole. This may be easily understood from the example of the fields of two ions acting upon three electrons with identical initial quasimomenta. We see from Fig. 59 that the changes in the quasimomenta of the three electrons are quite different because forces acting upon them depend on their position. Therefore *it is quite impossible to calculate the effect of internal forces $\mathbf{F}_d$ upon the distribution function from the laws of dynamics, and statistical laws should be invoked instead.*

External fields $\mathbf{F}$ result in "slow" changes in the state of the particle. Internal fields $\mathbf{F}_d$, on the other hand, may bring about sharp changes in the state of a particle in the short spell of time during which the electron passes through the small area of a located perturbation. Indeed, if the dimensions of the effective area of a localized perturbation do not exceed a few lattice constants, i.e. are of the order of 10^{-7} cm, and the particle velocity is around 10^7 cm/s (thermal velocity), the time of interaction with a localized centre will be only 10^{-14}s. *Such a brief interaction results in substantial changes in velocity and quasimomentum of the electron* which is equivalent to a collision of mechanical particles. Accordingly, it is termed *collision*. As a result of collisions the number of particles taking part in directional motion is changed, and for this reason *the term scattering is also applied to collision processes.*

To describe the different nature of external and internal forces we re-write (36.7) to take account of the two classes of forces, **F** and $\mathbf{F}_d$:

$$-\frac{\partial f}{\partial t}=(\mathbf{v},\ \nabla_{\mathbf{r}}f)+\frac{1}{\hbar}(\mathbf{F},\ \nabla_{\mathbf{k}}f)+\frac{1}{\hbar}(\mathbf{F}_d,\ \nabla_{\mathbf{k}}f). \tag{36.10}$$

The equation (36.10) shows that the distribution function changes are due to the motion of the particles with the velocity **v** and to the action of external **F** and internal $\mathbf{F}_d$ forces.

Denote the changes in the distribution function, brought about by the motion of the particles and by the action of external forces, by $\left(\frac{\partial f}{\partial t}\right)_f$. Evidently,

$$-\left(\frac{\partial f}{\partial t}\right)_f=(\mathbf{v},\ \nabla_{\mathbf{r}}f)+\frac{1}{\hbar}(\mathbf{F},\ \nabla_{\mathbf{k}}f), \tag{36.11}$$

$\left(\frac{\partial f}{\partial t}\right)_f$ is the *field term* of the Boltzmann equation. Denote the change of the distribution function caused by collisions by $\left(\frac{\partial f}{\partial t}\right)_c$:

$$-\left(\frac{\partial f}{\partial t}\right)_c=\frac{1}{\hbar}(\mathbf{F}_d,\ \nabla_{\mathbf{k}}f). \tag{36.12}$$

The quantity $\left(\frac{\partial f}{\partial t}\right)_c$ is termed the *collision integral* since it is represented in the form of an integral. Thus, we represent the variation of the distribution function in time as a sum of two terms—the field term $\left(\frac{\partial f}{\partial t}\right)_f$ and the collision integral $\left(\frac{\partial f}{\partial t}\right)_c$.

$$\frac{\partial f}{\partial t}=\left(\frac{\partial f}{\partial t}\right)_f+\left(\frac{\partial f}{\partial t}\right)_c \tag{36.13}$$

To find $\left(\frac{\partial f}{\partial t}\right)_c$ we will make use of statistical methods of description of physical processes.

Suppose that particles as a result of collisions change from the states (**r**, **k**) to the states (**r**′, **k**′). Let the probability of this transition per unit time be $w(\mathbf{k},\ \mathbf{k}')$. Obviously, the coordinate will not change greatly as a result of collision, therefore we may confine ourselves to the case $\mathbf{r}=\mathbf{r}'$ and neglect the dependence of w on **r**′ Take two elementary volumes $d\tau_{\mathbf{k}}$ and $d\tau_{\mathbf{k}'}$ around the points **k** and **k**′. The number of states in them with due account taken of spins is $\frac{d\tau_{\mathbf{k}}}{4\pi^3}$ and $\frac{d\tau_{\mathbf{k}'}}{4\pi^3}$ for a crystal of unit volume. The numbers of occupied states are $f(\mathbf{r},\ \mathbf{k})\frac{d\tau_{\mathbf{k}}}{4\pi^3}$ and $f(\mathbf{r},\ \mathbf{k}')\frac{d\tau_{\mathbf{k}'}}{4\pi^3}$, and of empty states $[1-f(\mathbf{r},\ \mathbf{k})]\frac{d\tau_{\mathbf{k}}}{4\pi^3}$ and $[1-f(\mathbf{r},\ \mathbf{k}')]\frac{d\tau_{\mathbf{k}'}}{4\pi^3}$, respectively. As a result of collisions the electrons move from $d\tau_{\mathbf{k}}$ to $d\tau_{\mathbf{k}'}$ and vice versa. The number of transitions will depend on the probability $w(\mathbf{k},\ \mathbf{k}')$ and also on the initial number of occupied states and final number of empty states (the Pauli principle should be applied). Therefore the change in the number of occupied states during the time dt as a result of direct electron transitions from $d\tau_{\mathbf{k}}$ to $d\tau_{\mathbf{k}'}$ and reverse transitions from $d\tau_{\mathbf{k}'}$ to $d\tau_{\mathbf{k}}$ will be equal to

$$-dt\left\{w(\mathbf{k},\ \mathbf{k}')\frac{d\tau_{\mathbf{k}}}{4\pi^3}f(\mathbf{r},\ \mathbf{k})[1-f(\mathbf{r},\ \mathbf{k}')]\frac{d\tau_{\mathbf{k}'}}{4\pi^3}\right\}+$$
$$+dt\left\{w(\mathbf{k}',\ \mathbf{k})\frac{d\tau_{\mathbf{k}'}}{4\pi^3}f(\mathbf{r},\ \mathbf{k}')[1-f(\mathbf{r},\ \mathbf{k})]\frac{d\tau_{\mathbf{k}}}{4\pi^3}\right\}. \qquad (36.14)$$

The first addend determines the reduction in the number of particles inside $d\tau_{\mathbf{k}}$ as a result of direct transitions from $d\tau_{\mathbf{k}}$ to $d\tau_{\mathbf{k}'}$, and *the second addend determines the increase in the number of particles inside $d\tau_{\mathbf{k}'}$ as a result of reverse transitions from $d\tau_{\mathbf{k}'}$ to $d\tau_{\mathbf{k}}$ with the probability* $w(\mathbf{k}',\ \mathbf{k})$. The total number of changes in the occupied states during the time dt may be found by integrating the expression (36.14) over the entire area of possible variations of $\mathbf{k}'$, i.e. over the Brillouin zone volume $V_{\mathbf{k}}$:

$$dt\frac{d\tau_{\mathbf{k}}}{4\pi^3}\int\limits_{(V_{\mathbf{k}})}\{w(\mathbf{k}',\ \mathbf{k})f(\mathbf{r},\ \mathbf{k}')[1-f(\mathbf{r},\ \mathbf{k})]-$$
$$-w(\mathbf{k},\ \mathbf{k}')f(\mathbf{r},\ \mathbf{k})[1-f(\mathbf{r},\ \mathbf{k}')]\}\frac{d\tau_{\mathbf{k}'}}{4\pi^3}. \qquad (36.15)$$

On the other hand, the number of occupied states in every moment of time is equal to $f(\mathbf{r},\ \mathbf{k})\frac{d\tau_{\mathbf{k}}}{4\pi^3}$ and its variation during time dt due to collisions may be expressed in the form

$$\left(\frac{\partial f}{dt}\right)_c dt\,\frac{d\tau_{\mathbf{k}}}{4\pi^3}. \qquad (36.16)$$

Equating the expressions (36.15) and (36.16) we obtain after cancelling out $dt\frac{d\tau_{\mathbf{k}}}{4\pi^3}$

$$\left(\frac{\partial f}{\partial t}\right)_c=-\frac{1}{\hbar}(\mathbf{F}_d,\ \nabla_{\mathbf{k}}f(\mathbf{r},\ \mathbf{k}))=$$
$$=\int\limits_{(V_{\mathbf{k}})}\{w(\mathbf{k}',\ \mathbf{k})f(\mathbf{r},\ \mathbf{k}')[1-f(\mathbf{r},\ \mathbf{k})]-w(\mathbf{k},\ \mathbf{k}')f(\mathbf{r},\ \mathbf{k})\times$$

$$\times[1-f(\mathbf{r},\ \mathbf{k}')]\}\frac{d\tau_{\mathbf{k}'}}{4\pi^3}=\int\limits_{(V_{\mathbf{k}})}\{[w(\mathbf{k}',\ \mathbf{k})f(\mathbf{r},\ \mathbf{k}')-$$

$$-w(\mathbf{k},\ \mathbf{k}')f(\mathbf{r},\ \mathbf{k}]-[w(\mathbf{k}',\ \mathbf{k})-w(\mathbf{k},\ \mathbf{k}')]f(\mathbf{r},\ \mathbf{k})f(\mathbf{r},\ \mathbf{k}')\}\frac{d\tau_{\mathbf{k}'}}{4\pi^3}. \tag{36.17}$$

If the probabilities of direct and reverse transitions are equal,

$$w(\mathbf{k},\ \mathbf{k}')=w(\mathbf{k}',\ \mathbf{k}), \tag{36.18}$$

the integral (36.17) may be simplified:

$$\left(\frac{\partial f}{\partial t}\right)_c=\int\limits_{(V_{\mathbf{k}})}w(\mathbf{k},\ \mathbf{k}')[f(\mathbf{r},\ \mathbf{k}')-f(\mathbf{r},\ \mathbf{k})]\frac{d\tau_{\mathbf{k}'}}{4\pi^3}. \tag{36.19}$$

Taking into account (36.13), (36.11) and (36.17) we obtain

$$\frac{\partial f}{\partial t}=-(\mathbf{v},\ \nabla_{\mathbf{r}}f)-\frac{1}{\hbar}(\mathbf{F},\ \nabla_{\mathbf{k}}f)+$$

$$+\int\limits_{(V_{\mathbf{k}})}w(\mathbf{k},\ \mathbf{k}')[f(\mathbf{r},\ \mathbf{k}')-f(\mathbf{r},\ \mathbf{k})]\frac{d\tau_{\mathbf{k}'}}{4\pi^3}. \tag{36.20}$$

The equation (36.20) is the so-called *Boltzmann kinetic equation*. Since the integral in (36.20) contains the function to be determined, the equation (36.20) is of the integro-differential type. As yet, there is no general solution to this equation.

For a stationary state $\frac{\partial f}{\partial t}=0$ the kinetic equation assumes the form

$$\left(\frac{\partial f}{\partial t}\right)_f=-\left(\frac{\partial f}{\partial t}\right)_c, \tag{36.21}$$

or

$$(\mathbf{v},\ \nabla_{\mathbf{r}}f(\mathbf{r},\ \mathbf{k}))+\frac{1}{\hbar}(\mathbf{F},\ \nabla_{\mathbf{k}}f(\mathbf{r},\ \mathbf{k}))=$$

$$=\int\limits_{(V_{\mathbf{k}})}w(\mathbf{k},\ \mathbf{k}')[f(\mathbf{r},\ \mathbf{k}')-f(\mathbf{r},\ \mathbf{k})]\frac{d\tau_{\mathbf{k}'}}{4\pi^3}. \tag{36.22}$$

It follows from expressions (36.21) or (36.22) that *in stationary conditions the changes in the distribution function brought about by the action of external fields and the motion of the particles are compensated by collisions of charge carriers with localized imperfections in the periodic structure of the lattice.* If $\left(\frac{\partial f}{\partial t}\right)_f\neq\left(\frac{\partial f}{\partial t}\right)_c$,

$\frac{\partial f}{\partial t} \neq 0$ and the distribution function will change with time, the nature of the change being dependent on which process predominates: the change in f due tó the action of the fields or as a result of charge carrier collisions.

The conditions for the kinetic equation to be applicable may be formulated as follows:

1. External action should not change the energy spectrum of the electron in the crystal. This limits the magnitude of the fields.

2. Since the kinetic equation is quasiclassical, it is not applicable to short-duration processes taking place inside small volumes since this leads to great indeterminacy in energy an^d quasimomentum.

Summary of Sec. 36

1. Physical phenomena brought about by the motion of charge carriers in a semiconductor in which a temperature gradient has been set up, or which is subjected to the action of external or internal fields are termed kinetic. The Boltzmann kinetic equation is used to describe them.

2. Since the total number of states in a crystal is a constant, the total time derivative of the distribution functions $f(\mathbf{r}, \mathbf{k}, t)$ is zero:

$$\frac{df}{dt} = 0. \tag{36.1s}$$

Differentiating $f(\mathbf{r}, \mathbf{k}, t)$ with respect to time as a complex function, i.e. regarding $\mathbf{r}$ and $\mathbf{k}$ as functions of time, and taking into account (36.1s) we obtain

$$-\frac{\partial f(\mathbf{r}, \mathbf{k}, t)}{\partial t} = (\mathbf{v}, \nabla_r f(\mathbf{r}, \mathbf{k}, t)) + \frac{1}{\hbar}(\mathbf{F}, \nabla_{\mathbf{k}} f(\mathbf{r}, \mathbf{k}, t)) \tag{36.2s}$$

The variation of the distribution function in time is represented by a sum of two terms—the field term and the collision term:

$$\frac{\partial f}{\partial t} = \left(\frac{\partial f}{\partial t}\right)_f + \left(\frac{\partial f}{\partial t}\right)_c, \tag{36.3s}$$

where

$$\left(\frac{\partial f}{\partial t}\right)_f = -(\mathbf{v}, \nabla_r f) - \frac{1}{\hbar}(\mathbf{F}, \nabla_{\mathbf{k}} f) \tag{36.4s}$$

and

$$\left(\frac{\partial f}{\partial t}\right)_c = -\frac{1}{\hbar}(\mathbf{F}_d, \nabla_{\mathbf{k}} f). \tag{36.5s}$$

3. As a result of the collisions the particles change their state with a probability $w(\mathbf{k}, \mathbf{k}')$. This enables the effect of the colli-

sions to be expressed in the form

$$\left(\frac{\partial f}{\partial t}\right)_c = \int\limits_{(V_{\mathbf{k}})} \{w(\mathbf{k}', \mathbf{k}) f(\mathbf{r}, \mathbf{k}') [1 - f(\mathbf{r}, \mathbf{k})] - \\ - w(\mathbf{k}, \mathbf{k}') f(\mathbf{r}, \mathbf{k}) [1 - f(\mathbf{r}, \mathbf{k}')]\} \frac{d\tau_{\mathbf{k}'}}{4\pi^3}. \quad (36.6s)$$

4. The Boltzmann kinetic equation is of the form:

$$\frac{\partial f(\mathbf{r}, \mathbf{k}, t)}{\partial t} = -(\mathbf{v}, \nabla_r f(\mathbf{r}, \mathbf{k}, t)) - \frac{1}{\hbar}(\mathbf{F}, \nabla_{\mathbf{k}} f(\mathbf{r}, \mathbf{k}, t)) + \left(\frac{\partial f}{\partial t}\right)_c. \quad (36.7s)$$

For a stationary state $w(\mathbf{k}, \mathbf{k}') = w(\mathbf{k}', \mathbf{k})$ the Boltzmann equation is

$$\left((\mathbf{v}, \nabla_r f(\mathbf{r}, \mathbf{k})) + \frac{1}{\hbar}(\mathbf{F}, \nabla_{\mathbf{k}} f(\mathbf{r}, \mathbf{k}))\right) = \\ = \frac{1}{4\pi^3} \int\limits_{(V_{\mathbf{k}})} w(\mathbf{k}, \mathbf{k}') [f(\mathbf{r}, \mathbf{k}') - f(\mathbf{r}, \mathbf{k})] d\tau_{\mathbf{k}'}. \quad (36.8s)$$

In the stationary state the changes brought about by the action of the field are compensated by collisions.

37. RELAXATION TIME

The general solution of the kinetic equation is a very intricate problem which may be substantially simplified if the so-called relaxation time may be introduced.

Suppose that at some moment of time $t = 0$ the field term is zero (the fields are switched off):

$$\left(\frac{\partial f}{\partial t}\right)_f = 0. \quad (37.1)$$

It follows from (36.13) that the distribution function will change as a result of collisions:

$$\frac{\partial f}{\partial t} = \left(\frac{\partial f}{\partial t}\right)_c. \quad (37.2)$$

If at the moment the fields were switched off the system was in a stationary non-equilibrium state, then after the fields are switched off the system should, as a result of collisions, return to the equilibrium state; in other words, *particle collisions restore equilibrium which was disturbed by the fields*. The simplest assumption concerning the relaxation process is that *the equilibrium restoration rate is proportional to the displacement* $[f(\mathbf{r}, \mathbf{k}, t) - f_0(\mathbf{r}, \mathbf{k})]$

from equilibrium:

$$\frac{\partial f}{\partial t}=\left(\frac{\partial f}{\partial t}\right)_c=-\frac{f(\mathbf{r},\ \mathbf{k},\ t)-f_0(\mathbf{r},\ \mathbf{k})}{\tau(\mathbf{k})}. \qquad (37.3)$$

f_0 in (37.3) is the distribution function for the state of equilibrium, $f(\mathbf{r},\ \mathbf{k},\ t)$ is the non-equilibrium distribution function, and $\frac{1}{\tau(\mathbf{k})}$ is a coefficient of proportionality dependent, generally, on $\mathbf{k}$ and $\mathbf{r}$. However, since below we shall be interested in the dependence of τ on $\mathbf{k}$, we shall neglect the dependence of τ on $\mathbf{r}$. Assuming τ to be positive we introduced a minus sign into (37.3) thereby providing for the return of the system to the equilibrium state. The solution of equation (37.3) is quite simple:

$$f(\mathbf{r},\ \mathbf{k},\ t)-f_0(\mathbf{r},\ \mathbf{k})=[f(\mathbf{r},\ \mathbf{k},\ 0)-f_0(\mathbf{r},\ \mathbf{k})]\,e^{-\frac{t}{\tau}}. \qquad (37.4)$$

The quantity $\tau(\mathbf{k})$ *shows the rate of the return to the state of equilibrium disturbed by external fields, therefore it is termed relaxation time*. The relaxation time enables the collision integral to be expressed in a more simple form. Indeed, it follows from (37.3) and (36.19) that

$$\left(\frac{\partial f}{\partial t}\right)_c=\int\limits_{(V_\mathbf{k})} w(\mathbf{k},\ \mathbf{k}')\,[f(\mathbf{r},\ \mathbf{k}',\ t)-f(\mathbf{r},\ \mathbf{k},\ t)]\,\frac{d\tau_{\mathbf{k}'}}{4\pi^3}=-\frac{f-f_0}{\tau(\mathbf{k})}. \qquad (37.5)$$

(37.5) leads to an "explicit" expression for the relaxation time in terms of scattering probability and distribution function:

$$\frac{1}{\tau(\mathbf{k})}=\frac{1}{4\pi^3}\int\limits_{(V_\mathbf{k})} w(\mathbf{k},\ \mathbf{k}')\,\frac{[f(\mathbf{r},\ \mathbf{k},\ t)-f(\mathbf{r},\ \mathbf{k}',\ t)]}{[f(\mathbf{r},\ \mathbf{k},\ t)-f_0(\mathbf{r},\ \mathbf{k})]}\,d\tau_{\mathbf{k}'}. \qquad (37.6)$$

We shall need an additional assumption to solve the Boltzmann equation: the relaxation time $\tau(\mathbf{k})$ is an unambiguous characteristic of the collision processes both during the relaxation process and the time external fields are active; in other words, we suppose that *the relaxation time is independent of external fields*. We shall discuss this assumption later. Now we shall simply use it to solve the Boltzmann equation which describes a stationary state. It follows from (36.8s) and (37.5) that

$$(\mathbf{v},\ \nabla_\mathbf{r} f(\mathbf{r},\ \mathbf{k}))+\frac{1}{\hbar}(\mathbf{F},\ \nabla_\mathbf{k} f(\mathbf{r},\ \mathbf{k}))=-\frac{f(\mathbf{r},\ \mathbf{k})-f_0(\mathbf{r},\ \mathbf{k})}{\tau(\mathbf{k})}. \qquad (37.7)$$

The expression (37.7) constitutes the principal equation which describes kinetic phenomena in a stationary state provided relaxation time may be introduced to take account of collisions. We will seek the solution of the equation (37.7) in the form of a series:

$$f(\mathbf{r},\mathbf{k}) = f_0(\mathbf{r},\mathbf{k}) + f^{(1)}(\mathbf{r},\mathbf{k}) + f^{(2)}(\mathbf{r},\mathbf{k}) + \ldots, \tag{37.8}$$

where $f^{(1)}(\mathbf{r},\mathbf{k})$; $f^{(2)}(\mathbf{r},\mathbf{k})$ are the first, second, etc. approximation corrections to the equilibrium distribution function. The **r** and **k** derivatives of $f^{(1)}(\mathbf{r},\mathbf{k})$ should be of the same order of magnitude as $f^{(2)}(\mathbf{r},\mathbf{k})$; therefore, to find the first approximation of (37.7), we may neglect the terms of higher orders. So we are seeking the solution of the equation (37.7) in the first approximation:

$$f(\mathbf{r},\mathbf{k}) = f_0(\mathbf{r},\mathbf{k}) + f^{(1)}(\mathbf{r},\mathbf{k}). \tag{37.9}$$

Substituting (37.9) into (37.7) we obtain the equation for $f^{(1)}(\mathbf{r},\mathbf{k})$:

$$(\mathbf{v},\ \nabla_r f_0(\mathbf{r},\mathbf{k}) + \nabla_r f^{(1)}(\mathbf{r},\mathbf{k})) + \frac{1}{\hbar}(\mathbf{F},\ \nabla_{\mathbf{k}} f_0(\mathbf{r},\mathbf{k}) + \\ + \nabla_{\mathbf{k}} f^{(1)}(\mathbf{r},\mathbf{k})) = -\frac{f^{(1)}(\mathbf{r},\mathbf{k})}{\tau(\mathbf{k})}. \tag{37.10}$$

Calculate $\nabla_r f_0(\mathbf{r},\mathbf{k})$ and $\nabla_{\mathbf{k}} f_0(\mathbf{r},\mathbf{k})$ taking into account that F and T may depend on **r**, and that E depends on **k**:

$$\nabla_r f_0(\mathbf{r},\mathbf{k}) = -\frac{e^{\frac{E-F}{kT}}\ \nabla_r\left(\frac{E-F}{kT}\right)}{\left[e^{\frac{E-F}{kT}}+1\right]^2} = -\frac{\partial f_0}{\partial E}\left\{\nabla_r F + (E-F)\frac{\nabla_r T}{T}\right\} \tag{37.11}$$

and

$$\nabla_{\mathbf{k}} f_0(\mathbf{r},\mathbf{k}) = -\frac{e^{\frac{E-F}{kT}}}{\left[e^{\frac{E-F}{kT}}+1\right]^2}\frac{\nabla_{\mathbf{k}} E}{kT} = \frac{\partial f_0}{\partial E}\hbar\mathbf{v}. \tag{37.12}$$

We see that the expressions for the gradients $\nabla_r f_0$ and $\nabla_{\mathbf{k}} f_0$ include the term

$$\frac{\partial f_0}{\partial E} = -\frac{e^{\frac{E-F}{kT}}}{\left[e^{\frac{E-F}{kT}}+1\right]^2}\frac{1}{kT}. \tag{37.13}$$

Making use of (37.11) and (37.12) reduce the equation (37.10) for $f^{(1)}$ to the form

$$-\frac{f^{(1)}(\mathbf{r},\mathbf{k})}{\tau(\mathbf{k})} = -(\mathbf{v},\ \nabla_r F + (E-F)\nabla_r \ln T)\frac{\partial f_0}{\partial E} + \\ + (\mathbf{v},\ \nabla_r f^{(1)}(\mathbf{r},\mathbf{k})) + (e\mathbf{E} + e[\mathbf{vB}],\ \mathbf{v})\frac{\partial f_0}{\partial E} + \frac{1}{\hbar}(e\mathbf{E} + e[\mathbf{vB}],\ \nabla_{\mathbf{k}} f^{(1)}), \tag{37.14}$$

where the expression for the Lorentz force was substituted for **F**.

Assuming that the derivatives of $f^{(1)}$ are of the order of magnitude comparable to $f^{(2)}$ and neglecting the second and the fourth terms in (37.14) we obtain

$$-\frac{f^{(1)}}{\tau(\mathbf{k})}=\frac{\partial f_0}{\partial E}(\{e\mathbf{E}-\nabla_{\mathbf{r}}F-(E-F)\nabla_{\mathbf{r}}\ln T\},\ \mathbf{v}). \qquad (37.15)$$

It follows from (37.15) that $f^{(1)}(\mathbf{r},\ \mathbf{k})$ to the first approximation is independent of the magnetic field. To find the dependence on $\mathbf{B}$ one should retain the term with $\nabla_{\mathbf{k}}f^{(1)}(\mathbf{r},\ \mathbf{k})$ in (37.14).

Calculate it using (37.15). First, however, write $f^{(1)}(\mathbf{r},\ \mathbf{k})$ making use of the expression (37.15), in the form

$$f^{(1)}(\mathbf{r},\ \mathbf{k})=-\frac{\partial f_0}{\partial E}(\mathbf{X}(\mathbf{r},\ \mathbf{k}),\ \mathbf{v}), \qquad (37.16)$$

where $\mathbf{X}(\mathbf{r},\ \mathbf{k})$ is an unknown vector function to which the problem of finding $f^{(1)}(\mathbf{r},\ \mathbf{k})$ has been reduced. It follows from comparison of (37.16) and (37.15) that neglecting the magnetic field we may write for $\mathbf{X}(\mathbf{r},\ \mathbf{k})$

$$\mathbf{X}(\mathbf{r},\ \mathbf{k})=\tau(\mathbf{k})\{e\mathbf{E}-\nabla_{\mathbf{r}}F-(E-F)\nabla_{\mathbf{r}}\ln T\}. \qquad (37.17)$$

Since $\mathbf{E}=-\nabla_{\mathbf{r}}\varphi(\mathbf{r})$, we obtain

$$\mathbf{X}(\mathbf{r},\ \mathbf{k})=-\tau(\mathbf{k})\{\nabla_{\mathbf{r}}(e\varphi+F)+(E-F)\nabla_{\mathbf{r}}\ln T\}. \qquad (37.18)$$

Now calculate $\nabla_{\mathbf{k}}f^{(1)}(\mathbf{r},\ \mathbf{k})$ taking into account (37.16):

$$\nabla_{\mathbf{k}}f^{(1)}(\mathbf{r},\ \mathbf{k})=-\frac{d}{d\mathbf{k}}\left\{\frac{\partial f_0}{\partial E}\left(\mathbf{X}\frac{dE}{\hbar\,d\mathbf{k}}\right)\right\}=-\frac{\partial f_0}{\partial E}\left(\mathbf{X}\frac{\hbar}{\mathbf{m}^*}\right)-$$
$$-\left(\mathbf{v},\ \frac{d}{d\mathbf{k}}\frac{\partial f_0}{\partial E}\mathbf{X}\right). \qquad (37.19)$$

$\frac{1}{\mathbf{m}^*}=\frac{d^2E}{\hbar^2d^2\mathbf{k}^2}$ in the expression (37.19) is the generalized inverse effective mass tensor. If (37.19) is substituted into (37.14) where only the term with $\mathbf{B}$ should be retained the last addend in (37.14) will yield

$$\frac{e}{\hbar}([\mathbf{vB}],\ \nabla_{\mathbf{k}}f^{(1)})=-\frac{e}{\hbar}\left([\mathbf{vB}],\ \left(\mathbf{X}\frac{\hbar}{\mathbf{m}^*}\right)\right)\frac{\partial f_0}{\partial E}-$$
$$-\frac{e}{\hbar}\left([\mathbf{vB}],\ \left(\mathbf{v},\ \nabla_{\mathbf{k}}\frac{\partial f_0}{\partial E}\mathbf{X}\right)\right)=e\left(\left[\frac{\mathbf{X}}{\mathbf{m}^*},\ \mathbf{B}\right];\ \mathbf{v}\right)\frac{\partial f_0}{\partial E}; \qquad (37.20)$$

the second addend turning zero because of $[\mathbf{vv}]=0$. Therefore, (37.14) will assume the form

$$-\frac{f^{(1)}(\mathbf{r},\ \mathbf{k})}{\tau(\mathbf{k})}=\frac{1}{\tau(\mathbf{k})}\frac{\partial f_0}{\partial E}(\mathbf{Xv})=$$
$$=\frac{\partial f_0}{\partial E}\left(\mathbf{v},\ e\mathbf{E}-\nabla_{\mathbf{r}}F-(E-F)\nabla_{\mathbf{r}}\ln T+e\left[\frac{\mathbf{X}}{\mathbf{m}^*},\ \mathbf{B}\right]\right). \qquad (37.21)$$

We obtained an equation for $\mathbf{X}(\mathbf{r}, \mathbf{k})$:

$$\mathbf{X}(\mathbf{r}, \mathbf{k}) = -\tau(\mathbf{k})\left\{\nabla_{\mathbf{r}}(F+e\varphi)+(E-F)\nabla_{\mathbf{r}}\ln T - e\left[\frac{\mathbf{X}}{\mathbf{m}^*}, \mathbf{B}\right]\right\}. \tag{37.22}$$

To solve it introduce brief notations

$$\nabla_{\mathbf{r}}(F+e\varphi)+(E-F)\nabla_{\mathbf{r}}\ln T = -\mathbf{L} \tag{37.23}$$

$$\tau(\mathbf{k})\,\mathbf{L} = \mathbf{A}. \tag{37.24}$$

In new notations the equation (37.22) will be

$$\mathbf{X} = \mathbf{A} + e\tau\left[\frac{\mathbf{X}}{\mathbf{m}^*}, \mathbf{B}\right]. \tag{37.25}$$

The solution of the equation (37.25) entails different procedure for scalar and tensor $\mathbf{m}^*$s.

1. Scalar effective mass. Since $\mathbf{m}^*$ now is a scalar, we may introduce a vector $\boldsymbol{\varphi}(\mathbf{k})$

$$\boldsymbol{\varphi}(\mathbf{k}) = \frac{e\tau(\mathbf{k})}{m^*}\mathbf{B}. \tag{37.26}$$

Now we may write (37.25) in the form

$$\mathbf{X} = \mathbf{A} + [\mathbf{X}\boldsymbol{\varphi}]. \tag{37.27}$$

The scalar product of (37.27) by $\boldsymbol{\varphi}$ is

$$(\mathbf{X}\boldsymbol{\varphi}) = (\mathbf{A}\boldsymbol{\varphi}) + ([\mathbf{X}\boldsymbol{\varphi}]\boldsymbol{\varphi}) = (\mathbf{A}\boldsymbol{\varphi}), \tag{37.28}$$

since

$$([\mathbf{X}\boldsymbol{\varphi}]\boldsymbol{\varphi}) = ([\boldsymbol{\varphi}\boldsymbol{\varphi}]\mathbf{X}) = 0. \tag{37.29}$$

The vector product of (37.27) by $\boldsymbol{\varphi}$ is

$$\begin{aligned}[\mathbf{X}\boldsymbol{\varphi}] &= [\mathbf{A}\boldsymbol{\varphi}] + [[\mathbf{X}\boldsymbol{\varphi}]\boldsymbol{\varphi}] = [\mathbf{A}\boldsymbol{\varphi}] + \boldsymbol{\varphi}(\mathbf{X}\boldsymbol{\varphi}) - \mathbf{X}(\boldsymbol{\varphi}\boldsymbol{\varphi}) = \\ &= [\mathbf{A}\boldsymbol{\varphi}] + \boldsymbol{\varphi}(\mathbf{A}\boldsymbol{\varphi}) - \mathbf{X}\boldsymbol{\varphi}^2 = \mathbf{X} - \mathbf{A}.\end{aligned} \tag{37.30}$$

When simplifying the expressions in the chain of equations (37.30) we made use of the rule for expanding the double vector product in the multipliers of a single vector product and replaced $(\mathbf{X}\boldsymbol{\varphi})$ and $[\mathbf{X}\boldsymbol{\varphi}]$ by $(\mathbf{A}\boldsymbol{\varphi})$ and $\mathbf{X}-\boldsymbol{\varphi}$ in accordance with (37.28) and (37.27). From (37.30) follow the expressions for $\mathbf{X}$:

$$\mathbf{X}(1+\boldsymbol{\varphi}^2) = \mathbf{A} + [\mathbf{A}\boldsymbol{\varphi}] + \boldsymbol{\varphi}(\mathbf{A}\boldsymbol{\varphi}), \tag{37.31}$$

$$\mathbf{X} = \frac{\mathbf{A} + [\mathbf{A}\boldsymbol{\varphi}] + \boldsymbol{\varphi}(\mathbf{A}\boldsymbol{\varphi})}{1+\boldsymbol{\varphi}^2}. \tag{37.32}$$

Substituting the expression for **A** and φ we obtain

$$\mathbf{X}(\mathbf{r}, \mathbf{k}) = \frac{-\tau\{\nabla_{\mathbf{r}}(e\varphi+F)+(E-F)\nabla_{\mathbf{r}}\ln T\}}{1+\frac{e^2\tau^2}{m^{*2}}B^2} \longrightarrow$$

$$\longrightarrow \frac{-\frac{e\tau^2}{m^*}[\nabla_{\mathbf{r}}(e\varphi+F)+(E-F)\nabla_{\mathbf{r}}\ln T, \mathbf{B}]}{1+\frac{e^2\tau^2}{m^{*2}}B^2} -$$

$$\longrightarrow \frac{-\frac{e^2\tau^3}{m^{*2}}\mathbf{B}(\nabla_{\mathbf{r}}(e\varphi+F)+(E-F)\nabla_{\mathbf{r}}\ln T, \mathbf{B})}{1+\frac{e^2\tau^2}{m^{*2}}B^2}. \tag{37.33}$$

The effects which depend on $f^{(1)}$ (**r**, **k**) *and, consequently, on* **X** (**r**, **k**) *are termed transverse if* $\mathbf{B} \perp \mathbf{L}$, *and longitudinal if* $\mathbf{B} \parallel \mathbf{L}$. *For transverse effects* $(\mathbf{L}, \mathbf{B}) = 0$, *and*

$$\mathbf{X} = \frac{\mathbf{A}+[\mathbf{A}\boldsymbol{\varphi}]}{1+\varphi^2}. \tag{37.34}$$

For longitudinal effects $[\mathbf{LB}] = 0$, *and*

$$\mathbf{X} = \frac{\mathbf{A}+\boldsymbol{\varphi}(\mathbf{A}\boldsymbol{\varphi})}{1+\varphi^2} = \frac{\mathbf{A}+\mathbf{A}\varphi^2}{1+\varphi^2} = \mathbf{A} =$$
$$= -\tau\{\nabla_{\mathbf{r}}(e\varphi+F)+(E-F)\nabla_{\mathbf{r}}\ln T\}, \tag{37.35}$$

i.e. if m^* *is a scalar all phenomena take the same course in a longitudinal magnetic field as when* $\mathbf{B} = 0$. *The presence of longitudinal effects points to the tensor nature of the effective mass.*

2. **Tensor effective mass.** We suppose that $\mathbf{m}^{*-1}$ is of diagonal type. In this case $\mathbf{m}^*$ will be diagonal too. Consider now the vector product $[\mathbf{m}^{*-1}\mathbf{X}, \mathbf{B}] = \mathbf{D}$. Denote $\mathbf{m}^{*-1}\mathbf{X} = \boldsymbol{\eta}$. It is easily seen that

$$\mathbf{X} = \mathbf{m}^*\boldsymbol{\eta}. \tag{37.36}$$

Find the components of the vector **D**:

$$D_x = \eta_y B_z - \eta_z B_y; \quad D_y = \eta_z B_x - \eta_x B_z;$$
$$D_z = \eta_x B_y - \eta_y B_x. \tag{37.37}$$

But

$$\eta_i = \{\mathbf{m}^{*-1}\mathbf{X}\}_i = \sum_l \mathbf{m}_{il}^{*-1} X_l = m_i^{-1} X_i = \frac{X_i}{m_i} \tag{37.38}$$

and for this reason D_x, for example, assumes the form

$$D_x = \frac{X_y B_z}{m_y} - \frac{X_z B_y}{m_z}. \tag{37.39}$$

Consider $\mathbf{m}^{*-1}\mathbf{D}$ determined by its components:

$$\{\mathbf{m}^{*-1}\mathbf{D}\}_i \doteq \sum_j \mathbf{m}^{*-1}_{ij} D_j = \frac{D_i}{m_i}, \tag{37.40}$$

or, taking account of (37.37),

$$\begin{aligned}\{\mathbf{m}^{*-1}\mathbf{D}\}_x &= \frac{D_x}{m_x} = \frac{X_y B_z}{m_x m_y} - \frac{X_z B_y}{m_x m_z} = \\ &= \frac{1}{m_x m_y m_z}(X_y m_z B_z - X_z m_y B_y) = \\ &= |\mathbf{m}^{*-1}|\Big(X_y \sum_i m_{zi} B_i - X_z \sum_i m_{yi} B_i\Big),\end{aligned} \tag{37.41}$$

i.e.

$$\mathbf{m}^{*-1}\mathbf{D} = \mathbf{m}^{*-1}[\mathbf{m}^{*-1}\mathbf{X},\ \mathbf{B}] = |\mathbf{m}^{*-1}|[\mathbf{X},\ \mathbf{m}^*\mathbf{B}]. \tag{37.42}$$

$|\mathbf{m}^*|$ denotes the determinant of tensor $\mathbf{m}^*$ matrix

$$|\mathbf{m}^*| = \mathrm{Det}\begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{pmatrix} = m_1 m_2 m_3. \tag{37.43}$$

Consider again the equation (37.25) for $\mathbf{X}$. Multiply it scalarly by $\mathbf{m}^{*-1}$ to obtain, with the account of (37.42),

$$\begin{aligned}\mathbf{m}^{*-1}\cdot\mathbf{X} &= \mathbf{m}^{*-1}\cdot\mathbf{A} + e\tau\mathbf{m}^{*-1}[\mathbf{m}^{*-1}\mathbf{X},\ \mathbf{B}] = \\ &= \mathbf{m}^{*-1}\cdot\mathbf{A} + \frac{e\tau}{|\mathbf{m}^*|}[\mathbf{X},\ \mathbf{m}^*\mathbf{B}].\end{aligned} \tag{37.44}$$

Multiplying (37.44) vectorially by $\mathbf{B}$ we obtain

$$[\mathbf{m}^{*-1}\mathbf{X},\ \mathbf{B}] = [\mathbf{m}^{*-1}\mathbf{A},\ \mathbf{B}] + \frac{e\tau}{|\mathbf{m}^*|}[[\mathbf{X},\ \mathbf{m}^*\mathbf{B}]\,\mathbf{B}]. \tag{37.45}$$

Decompose the double vector product

$$[[\mathbf{X},\ \mathbf{m}^*\mathbf{B}]\,\mathbf{B}] = (\mathbf{XB})\,\mathbf{m}^*\mathbf{B} - (\mathbf{B},\ \mathbf{m}^*\mathbf{B})\,\mathbf{X} \tag{37.46}$$

and substitute it into (37.45) to obtain

$$[\mathbf{m}^{*-1}\mathbf{X},\ \mathbf{B}] = [\mathbf{m}^{*-1}\mathbf{A},\ \mathbf{B}] + \frac{e\tau}{|\mathbf{m}^*|}\mathbf{m}^*\cdot\mathbf{B}\,(\mathbf{XB}) - \frac{e\tau}{|\mathbf{m}^*|}(\mathbf{B},\ \mathbf{m}^*\mathbf{B})\,\mathbf{X}. \tag{37.47}$$

It follows, however, from (37.25) that

$$(\mathbf{XB}) = (\mathbf{AB}), \tag{37.48}$$

and

$$e\tau[\mathbf{m}^{*-1}\mathbf{X},\ \mathbf{B}] = \mathbf{X} - \mathbf{A}. \tag{37.49}$$

Therefore, comparing (37.49) and (37.47), we write

$$\mathbf{X} - \mathbf{A} = e\tau[\mathbf{m}^{*-1}\mathbf{A},\ \mathbf{B}] + \frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{AB})\,\mathbf{m}^*\cdot\mathbf{B} - \frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\ \mathbf{m}^*\mathbf{B})\,\mathbf{X}. \tag{37.50}$$

Collecting similar terms we obtain

$$\mathbf{X}=\frac{\mathbf{A}+e\tau\left[\mathbf{m}^{*-1}\mathbf{A},\ \mathbf{B}\right]+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{AB})\,\mathbf{m}^*\cdot\mathbf{B}}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\ \mathbf{m}^*\mathbf{B})}. \qquad (37.51)$$

If the effective mass is a scalar, i.e. if all the diagonal elements are equal to m^*, $|\mathbf{m}^*|=m^{*3}$, and the equation (37.51) will turn into the equation (37.32).

Summary of Sec. 37

1. Collision processes restore equilibrium of electron and hole distribution disturbed by the action of the fields. The effect of collision processes may be described with the aid of relaxation time $\tau(\mathbf{k})$ defined by the collision integral of expression (37.6). The relaxation time is equal to the time non-equilibrium state persists, on the average, after the fields responsible for it have been switched off.

2. The Boltzmann equation for stationary states in the assumption that relaxation time is independent of external fields is

$$(\mathbf{v},\ \nabla_{\mathbf{r}} f(\mathbf{r},\ \mathbf{k}))+\frac{1}{\hbar}(\mathbf{F},\ \nabla_{\mathbf{k}} f(\mathbf{r},\ \mathbf{k}))=-\frac{f(\mathbf{r},\ \mathbf{k})-f_0(\mathbf{r},\ \mathbf{k})}{\tau(\mathbf{k})}. \qquad (37.1\text{s})$$

3. Solving the equation (37.1s) with the aid of the method of expanding it into a series of varying field magnitudes we obtain the equation for the first approximation correction:

$$-\frac{f^{(1)}(\mathbf{r},\ \mathbf{k})}{\tau(\mathbf{k})}=\frac{\partial f_0}{\partial E}(\mathbf{v},\ e\mathbf{E}+\nabla_{\mathbf{r}}(E-F)-(E-F)\nabla_{\mathbf{r}}\ln T)+$$
$$+\frac{e}{\hbar}([\mathbf{vB}]\,\nabla_{\mathbf{k}} f^{(1)}). \qquad (37.2\text{s})$$

4. Having written $f^{(1)}(\mathbf{r},\ \mathbf{k})$ in the form

$$f^{(1)}(\mathbf{r},\ \mathbf{k})=-\frac{\partial f_0}{\partial E}(\mathbf{X}(\mathbf{r},\ \mathbf{k}),\ \mathbf{v}) \qquad (37.3\text{s})$$

we obtain the equation for $\mathbf{X}(\mathbf{r},\ \mathbf{k})$

$$\mathbf{X}(\mathbf{r},\ \mathbf{k})=\mathbf{A}+e\tau\left[\frac{\mathbf{X}}{\mathbf{m}^*},\ \mathbf{B}\right], \qquad (37.4\text{s})$$

where

$$\mathbf{A}=\tau\mathbf{L}=-\tau(\mathbf{k})\left\{\nabla_{\mathbf{r}}(e\varphi+F)+(E-F)\nabla_{\mathbf{r}}\ln T\right\}. \qquad (37.5\text{s})$$

5. The solution for $\mathbf{X}(\mathbf{r}, \mathbf{k})$ in the case of the tensor effective mass is of the form

$$\mathbf{X}(\mathbf{r}, \mathbf{k}) = \frac{\mathbf{A} + e\tau[\mathbf{m}^{*-1} \cdot \mathbf{A}, \mathbf{B}] + \frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{A}b)\,\mathbf{m}^*\mathbf{B}}{1 + \frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B}, \mathbf{m}^*\mathbf{B})}. \tag{37.6s}$$

In the case of a scalar effective mass

$$\mathbf{X}(\mathbf{r}, \mathbf{k}) = \frac{\mathbf{A} + \frac{e\tau}{m^*}[\mathbf{AB}] + \frac{e^2\tau^2}{m^{*2}}(\mathbf{AB})\,\mathbf{B}}{1 + \frac{e^2\tau^2}{m^{*2}}\mathbf{B}^2}. \tag{37.7s}$$

Substituting (37.7s) or (37.6s) into (37.3s) and taking account of (37.5s) we obtain the solution of the Boltzmann equation for a stationary state.

6. All equations obtained in this section are equally valid both for electrons and holes provided the parameters of the respective particles are substituted into the equations. This applies not only to the charge e, effective mass $\mathbf{m}^*$ and relaxation time τ, but to the distribution function $f_0(\mathbf{r}, \mathbf{k})$ subject to conditions stated in Sec. 25, as well. If the semiconductor contains charge carriers of different types, each type may be described by its own distribution function.

38. ELECTRIC CURRENT DENSITY AND ENERGY FLUX DENSITY

As was repeatedly stated above, the prerequisite for a directional charge carrier current is the distortion of the distribution function symmetry, i.e. the disturbance of the charge system equilibrium. There are $2\frac{d\tau_\mathbf{k}}{8\pi^3}$ states in an element of volume $d\tau_\mathbf{k}$ in a crystal of unit volume. The number of charge carriers (electrons or holes) in these states is

$$dn = \frac{d\tau_\mathbf{k}}{4\pi^3} f(\mathbf{r}, \mathbf{k}). \tag{38.1}$$

Their velocity is $\mathbf{v} = \frac{1}{\hbar}\frac{dE}{d\mathbf{k}}$, and the elementary current density

$$d\mathbf{j} = e\mathbf{v}\frac{d\tau_\mathbf{k}}{4\pi^3} f(\mathbf{r}, \mathbf{k}). \tag{38.2}$$

The total current density is

$$\mathbf{j} = \frac{e}{4\pi^3}\int\limits_{(V_\mathbf{k})} \mathbf{v} f(\mathbf{r}, \mathbf{k})\, d\tau_\mathbf{k} = \frac{e}{4\pi^3}\int\limits_{(V_\mathbf{k})} \mathbf{v}(\mathbf{k}) f^{(1)}(\mathbf{r}, \mathbf{k})\, d\tau_\mathbf{k}. \tag{38.3}$$

In (38.3) we took into account that

$$\int\limits_{(V_{\mathbf{k}})} \mathbf{v} f_0(\mathbf{r}, \mathbf{k})\, d\tau_{\mathbf{k}} = 0, \tag{38.4}$$

because $f_0(\mathbf{r}, \mathbf{k})$ is an even function of $\mathbf{k}$, and $\mathbf{v}f_0(\mathbf{r}, \mathbf{k})$, an odd function which, when integrated within symmetrical limits, yields zero. In terms of physics, this means that *there is no current in a substance in conditions of thermodynamical equilibrium.* $\mathbf{j}$ may thus be expressed in the form

$$\mathbf{j} = -\frac{e}{4\pi^3} \int\limits_{(V_{\mathbf{k}})} \mathbf{v} \frac{\partial f_0}{\partial E} (\mathbf{X}\mathbf{v})\, d\tau_{\mathbf{k}} = -\frac{e}{4\pi^3 \hbar^2} \int\limits_{(V_{\mathbf{k}})} \frac{\partial f_0}{\partial E} \left(\mathbf{X} \frac{dE}{d\mathbf{k}} \right) \frac{dE}{d\mathbf{k}} \cdot d\tau_{\mathbf{k}}. \tag{38.5}$$

Directional flow of charged particles not only establishes electric current but, since each particle carries the energy E, causes energy to be transported, as well. *The energy flux density* $\mathbf{W}$, *i.e.*, *the amount of energy passing through cross section of unit area normal to the direction of flow per unit time is given by the expression*

$$\mathbf{W} = \frac{1}{4\pi^3} \int\limits_{(V_{\mathbf{k}})} E\mathbf{v} f^{(1)}(\mathbf{r}, \mathbf{k})\, d\tau_{\mathbf{k}} = -\frac{1}{4\pi^3} \int\limits_{(V_{\mathbf{k}})} E\mathbf{v} \frac{\partial f_0}{\partial E} (\mathbf{X}\mathbf{v})\, d\tau_{\mathbf{k}}. \tag{38 6}$$

If (37.51) is substituted for $\mathbf{X}$ into (38.5) and (38.6) we will obtain the expressions for $\mathbf{j}$ and $\mathbf{W}$:

$$\mathbf{j} = -\frac{e}{4\pi^3} \int\limits_{(V_{\mathbf{k}})} \frac{\partial f_0}{\partial E} \frac{\left(\left\{ \tau\mathbf{L} + e\tau^2 [\mathbf{m}^{*-1}\mathbf{L}, \mathbf{B}] + \frac{e^2\tau^3}{|\mathbf{m}^*|} (\mathbf{L}\mathbf{B})\, \mathbf{m}^*\mathbf{B} \right\} \mathbf{v} \right) \mathbf{v}\, d\tau_{\mathbf{k}}}{1 + \frac{e^2\tau^2}{|\mathbf{m}^*|} (\mathbf{B}, \mathbf{m}^*\mathbf{B})}, \tag{38.7}$$

$$\mathbf{W} = -\frac{1}{4\pi^3} \int\limits_{(V_{\mathbf{k}})} \frac{\partial f_0}{\partial E} E \frac{\left(\left\{ \tau\mathbf{L} + e\tau^2 [\mathbf{m}^{*-1}\mathbf{L}, \mathbf{B}] + \frac{e^2\tau^3}{|\mathbf{m}^*|} (\mathbf{L}\mathbf{B})\, \mathbf{m}^*\mathbf{B} \right\} \mathbf{v} \right) \mathbf{v}\, d\tau_{\mathbf{k}}}{1 + \frac{e^2\tau^2}{|\mathbf{m}^*|} (\mathbf{B}, \mathbf{m}^*\mathbf{B})}. \tag{33.8}$$

Before we calculate the expressions (38.7) and (38.8) consider the relation of the form

$$\mathbf{M} = -\frac{1}{4\pi^3} \int\limits_{(V_{\mathbf{k}})} E^{r-1} \tau^s \frac{\partial f_0}{\partial E} \frac{(\mathbf{G}\mathbf{v})\, \mathbf{v}\, d\tau_{\mathbf{k}}}{1 + \frac{e^2\tau^2}{|\mathbf{m}^*|} (\mathbf{B}, \mathbf{m}^*\mathbf{B})}, \tag{38.9}$$

where vector $\mathbf{G}$ is independent of $\mathbf{k}$. For the ith component M_i we may write

$$M_i = -\frac{1}{4\pi^3}\int\limits_{(V_\mathbf{k})} \frac{E^{r-1}\tau^s}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B}, \mathbf{m}^*\mathbf{B})} \frac{\partial f_0}{\partial E}\left(\sum_j G_j v_j\right) v_i \, d\tau_\mathbf{k}. \qquad (38.10)$$

If we denote

$$-\frac{1}{4\pi^3}\int\limits_{(V_\mathbf{k})} \frac{E^{r-1}\tau^s}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B}, \mathbf{m}^*\mathbf{B})} \frac{\partial f_0}{\partial E} v_i v_j \, d\tau_\mathbf{k} = K'^{ij}_{rs}, \qquad (38.11)$$

we may write (38.10) in the form

$$M_i = \sum_j K'^{ij}_{rs} G_j \qquad (38.12)$$

and

$$\mathbf{M} = K'_{rs}\mathbf{G}, \qquad (38.13)$$

where K'_{rs} is a second rank tensor, and K'^{ij}_{rs}—its ijth element given by the expression (38.11).

It follows from (38.7) and (38.8) that $\mathbf{j}$ and $\mathbf{W}$ may be expressed in terms of the K'_{rs}-type tensor to which we will assign the name of *(generalized) kinetic coefficient tensor.* Write $\mathbf{j}$ and $\mathbf{W}$ for the case of a scalar effective mass. Since $\mathbf{L}$ may be represented in the form

$$\begin{aligned}\mathbf{L} &= -\nabla_\mathbf{r}(e\varphi + F) - (E-F)\nabla_\mathbf{r}\ln T = \\ &= e\mathbf{E} - \nabla_\mathbf{r}F - E\frac{\nabla_\mathbf{r}T}{T} + \frac{E}{T}\nabla_\mathbf{r}T = e\mathbf{E} - E\frac{\nabla_\mathbf{r}T}{T} - T\nabla_\mathbf{r}\frac{F}{T},\end{aligned} \qquad (38.14)$$

we write, omitting the subscript $\mathbf{r}$ from the nabla operator,

$$\begin{aligned}\mathbf{j} = &-\frac{e}{12\pi^3}\int\limits_{(V_\mathbf{k})} \frac{\partial f_0}{\partial E} \frac{\tau v^2\left\{e\mathbf{E} - T\nabla\frac{F}{T} - E\frac{\nabla T}{T}\right\}}{1+\frac{e^2\tau^2}{m^{*2}}B^2} d\tau_\mathbf{k} - \\ &-\frac{e}{12\pi^3}\int\limits_{(V_\mathbf{k})} \frac{\partial f_0}{\partial E} \frac{\tau^2 v^2 \frac{e}{m^*}\left[e\mathbf{E} - T\nabla\frac{F}{T} - E\frac{\nabla T}{T}, \mathbf{B}\right] d\tau_\mathbf{k}}{1+\frac{e^2\tau^2}{m^{*2}}B^2} - \\ &-\frac{e}{12\pi^3}\int\limits_{(V_\mathbf{k})} \frac{\partial f_0}{\partial E} \frac{\tau^3 v^2 \frac{e^2}{m^{*2}}\mathbf{B}\left(e\mathbf{E} - T\nabla\frac{F}{T} - E\frac{\nabla T}{T}, \mathbf{B}\right) d\tau_\mathbf{k}}{1+\frac{e^2\tau^2}{m^{*2}}B^2}\end{aligned} \qquad (38.15)$$

and

$$\mathbf{W} = -\frac{1}{12\pi^3} \int\limits_{(V_{\mathbf{k}})} \frac{\partial f_0}{\partial E} \frac{E\tau v^2 \; ' \left. e\mathbf{E} - T\nabla \frac{F}{T} - E \frac{\nabla T}{T} \right\} d\tau_{\mathbf{k}}}{1 + \frac{e^2\tau^2}{m^{*2}} B^2} -$$

$$-\frac{1}{12\pi^3} \int\limits_{(V_{\mathbf{k}})} \frac{\partial f_0}{\partial E} \frac{E\tau^2 v^2 \frac{e}{m^*} \left[e\mathbf{E} - T\nabla \frac{F}{T} - E \frac{\nabla T}{T}, \ \mathbf{B} \right] d\tau_{\mathbf{k}}}{1 + \frac{e^2\tau^2}{m^{*2}} B^2} -$$

$$-\frac{1}{12\pi^3} \int\limits_{(V_{\mathbf{k}})} \frac{\partial f_0}{\partial E} \frac{E\tau^3 v^2 \frac{e^2}{m^{*2}} \mathbf{B} \left(e\mathbf{E} - T\nabla \frac{F}{T} - E \frac{\nabla T}{T}, \ \mathbf{B} \right) d\tau_{\mathbf{k}}}{1 + \frac{e^2\tau^2}{m^{*2}} B^2}. \qquad (38.16)$$

The relations (38.15) and (38.16) may be expressed in terms of the kinetic coefficients:

$$\mathbf{j} = \left(e\mathbf{E} - T\nabla \frac{F}{T} \right) eK'_{11} - e\frac{\nabla T}{T} K'_{21} + \frac{e^2}{m^*} \left[e\mathbf{E} - T\nabla \frac{F}{T}, \mathbf{B} \right] K'_{12} -$$
$$-\frac{e^2}{m^*} \left[\frac{\nabla T}{T}, \ \mathbf{B} \right] K'_{22} + \frac{e^3}{m^{*2}} \mathbf{B} \left(e\mathbf{E} - T\nabla \frac{F}{T}, \ \mathbf{B} \right) K'_{13} -$$
$$-\frac{e^3}{m^{*2}} \mathbf{B} \left(\frac{\nabla T}{T}, \ \mathbf{B} \right) K'_{23}. \qquad (38.17)$$

The expression for **W** is of a similar form, only the first index of the kinetic coefficient is greater by a unity, and the power of the charge—smaller by a unity:

$$\mathbf{W} = \left(e\mathbf{E} - T\nabla \frac{F}{T} \right) K'_{21} - \frac{\nabla T}{T} K'_{31} + \frac{e}{m^*} \left[e\mathbf{E} - T\nabla \frac{F}{T}, \ \mathbf{B} \right] K'_{22} -$$
$$-\frac{e}{m^*} \left[\frac{\nabla T}{T}, \ \mathbf{B} \right] K'_{32} + \frac{e^2}{m^{*2}} \mathbf{B} \left(e\mathbf{E} - T\nabla \frac{F}{T}, \ \mathbf{B} \right) K'_{23} -$$
$$-\frac{e^2}{m^{*2}} \mathbf{B} \left(\frac{\nabla T}{T}, \ \mathbf{B} \right) K'_{33}. \qquad (38.18)$$

Re-write the expression (38.17) for **j** collecting the terms:

$$\mathbf{j} = \left\{ e^2 K'_{11}\mathbf{E} - eK'_{11} T\nabla \frac{F}{T} - eK'_{21} \frac{\nabla T}{T} \right\} +$$
$$+ \left[\frac{e^3}{m^*} K'_{12}\mathbf{E} - \frac{e^2}{m^*} K'_{12} T\nabla \frac{F}{T} - \frac{e^2}{m^*} K'_{22} \frac{\nabla T}{T}, \ \mathbf{B} \right] +$$
$$+ \left(\frac{e^4}{m^{*2}} K'_{13}\mathbf{E} - \frac{e^3}{m^{*2}} K'_{13} T\nabla \frac{F}{T} - \frac{e^3}{m^{*2}} K'_{23} \frac{\nabla T}{T}; \ \mathbf{B} \right) \mathbf{B}. \qquad (38.19)$$

The first term in (38.19) *determines the conductivity current, the second and the third terms—the currents established by the gradients of the chemical potential and of temperature, i.e. these terms are related to diffusional and thermoelectric currents, the fourth* *term is responsible for transverse galvano- and thermomagnetic currents, the fifth term determines the changes in longitudinal currents due to the magnetic field.*

Summary of Sec. 38

1. Current density j and energy flux **W** are determined by the gradients of electrostatic potential φ, of chemical potential F, of temperature T, by the magnetic field **B** and by the properties of the substance described by the effective mass and the kinetic coefficient tensor K'_{rs}:

$$\mathbf{j}=\left\{e^2K'_{11}\mathbf{E}-eK'_{11}T\nabla\frac{F}{T}-eK'_{21}\frac{\nabla T}{T}\right\}+$$
$$+\left[\frac{e^3}{m^*}K'_{12}\mathbf{E}-\frac{e^2}{m^*}K'_{12}T\nabla\frac{F}{T}-\frac{e^2}{m^*}K'_{22}\frac{\nabla T}{T},\ \mathbf{B}\right]+$$
$$+\left(\frac{e^4}{m^{*2}}K'_{13}\mathbf{E}-\frac{e^3}{m^{*2}}K'_{13}T\nabla\frac{F}{T}-\frac{e^3}{m^{*2}}K'_{23}\frac{\nabla T}{T},\ \mathbf{B}\right)\mathbf{B} \qquad (38.1s)$$

and

$$\mathbf{W}=\left\{eK'_{21}\mathbf{E}-K'_{21}T\nabla\frac{F}{T}-K'_{31}\frac{\nabla T}{T}\right\}+$$
$$+\left[\frac{e^2}{m^*}K'_{22}\mathbf{E}-\frac{e}{m^*}K'_{22}T\nabla\frac{F}{T}-\frac{e}{m^*}K'_{32}\frac{\nabla T}{T},\ \mathbf{B}\right]+$$
$$+\left(\frac{e^3}{m^{*2}}K'_{23}\mathbf{E}-\frac{e^2}{m^{*2}}K'_{23}T\nabla\frac{F}{T}-\frac{e^2}{m^{*2}}K'_{33}\frac{\nabla T}{T},\ \mathbf{B}\right)\mathbf{B}. \qquad (38.2s)$$

2. The kinetic coefficient tensor K'_{rs} is determined by its elements K'^{ij}_{rs}:

$$K'^{ij}_{rs}=-\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})}\frac{E^{r-1}\tau^s}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\ \mathbf{m}^*\mathbf{B})}\frac{\partial f_0}{\partial E}v_iv_j\,d\tau_{\mathbf{k}}. \qquad (38.3s)$$

The physical meaning of K'_{rs} may be different depending on the subscripts r and s.

39. KINETIC COEFFICIENTS

Scalar effective mass. Since j and **W** are expressed in terms of tensors of kinetic coefficients K'_{rs}, we should be able to calculate them for different cases. Suppose the effective mass is a scalar. In this case, from considerations of symmetry, we may presume relaxation time to be isotropic: $\tau(\mathbf{k})=\tau(|\mathbf{k}|)$; in other words, *relaxation time should depend only on energy*, $\tau=\tau(E)$. Represent energy as a quadratic function of quasimomentum at every point of the Brillouin zone, taking into account that the expression for K'_{rs} includes $\frac{\partial f_0}{\partial E}$ which decreases rapidly with the increase in energy. The velocity may be expressed in terms of $\mathbf{k}$, or $\mathbf{P}$, with the aid of the usual relation

$$\mathbf{v}=\frac{\mathbf{P}}{m^*}=\frac{\hbar\mathbf{k}}{m^*} \tag{39.1}$$

and

$$E=E_0+\frac{\hbar^2\mathbf{k}^2}{2m^*}=E_0+\frac{m^*\mathbf{v}^2}{2}. \tag{39.2}$$

Write the expression for K'^{ij}_{rs}

$$K'^{ij}_{rs}=-\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})}\frac{E^{r-1}\tau^s}{1+\frac{e^2\tau^2}{m^{*2}}B^2}\frac{\partial f_0}{\partial E}v_iv_j\,d\tau_{\mathbf{k}}. \tag{39.3}$$

We reduce integration over the volume of the Brillouin zone to integration over energy. To this end express $d\tau_{\mathbf{k}}$ in terms of energy $dE(\mathbf{k})$:

$$dE=(\nabla_{\mathbf{k}}E,\ d\mathbf{k})=|\nabla_{\mathbf{k}}E|\,dk_n=\hbar\,|\mathbf{v}|\,dk_n, \tag{39.4}$$

where dk_n is the projection of the vector $d\mathbf{k}$ on the direction of the normal to the constant-energy surface. Introduce the element of the energy surface dS_E to express the element of volume in the form

$$d\tau_{\mathbf{k}}=dS_E\,dk_n=\frac{dS_E\,dF}{\hbar v}. \tag{39.5}$$

Substitute (39.5) into (39.3):

$$K'^{ij}_{rs}=-\frac{1}{4\pi^3}\int\limits_{(E)}\frac{E^{r-1}\tau^s}{1+\mu^2B^2}\frac{\partial f_0}{\partial E}dE\int\limits_{(S_E)}\frac{v_iv_j}{\hbar v}dS_E. \tag{39.6}$$

For the sake of simplification the notation was introduced into (39.6)

$$\mu=\frac{e\tau(E)}{m^*}=\mu(E), \tag{39.7}$$

where $\mu(E)$ is the mobility of charge carriers having the energy E. The integral over a surface may be represented as an integral over a space angle. Since

$$dS_E=k^2\,d\Omega, \tag{39.8}$$

it follows

$$\int\limits_{(S_E)}\frac{v_iv_j}{\hbar v}dS_E=M\int\limits_{(S_{1E})}\frac{\hbar k_i\hbar k_j}{m^{*2}\hbar^2\frac{k}{m^*}}k^2\,d\Omega=$$

$$=\frac{M}{m^*}k^3\int\limits_{(S_{1E})}\frac{k_ik_j}{k^2}d\Omega. \tag{39.9}$$

In (39.9) integration over the complete constant-energy surface was reduced to integration over one of the M energy surfaces S_{1E}. For a spherical coordinate system with the polar axis coinciding with k_z-direction in the Brillouin zone we write

$$\frac{k_x}{k}=\sin\theta\cos\varphi;\quad \frac{k_y}{k}=\sin\theta\sin\varphi;\quad \frac{k_z}{k}=\cos\theta. \tag{39.10}$$

Calculate (39.9) for two cases:

$$\int\limits_{(S_{1E})}\frac{k_xk_x}{k^2}d\Omega=\int\limits_0^{\pi}\int\limits_0^{2\pi}\sin^2\theta\cos^2\varphi\sin\theta\,d\theta\,d\varphi=\frac{4\pi}{3}, \tag{39.11}$$

$$\int\limits_{(S_{1E})}\frac{k_xk_y}{k^2}d\Omega=\int\limits_0^{\pi}\int\limits_0^{2\pi}\sin^2\theta\cos\varphi\sin\varphi\sin\theta\,d\theta\,d\varphi=0. \tag{39.12}$$

Making similar calculations we obtain, generally,

$$\int\limits_{(S_{1E})}\frac{k_ik_j}{k^2}d\Omega=\frac{4\pi}{3}\delta_{ij}, \tag{39.13}$$

i.e. *in case of spherical constant-energy surfaces the tensor* K'_{rs} *is diagonal.* Express the integral over the energy surface in terms of the value of energy on this surface:

$$\begin{aligned}\int\limits_{(S)}\frac{v_iv_j}{\hbar v}dS_E&=\frac{M}{m^*}k^3\frac{4\pi}{3}\delta_{ij}=\\&=\frac{4\pi}{3}\frac{M}{m^*}\left[\frac{2m^*(E-E_0)}{\hbar^2}\right]^{3/2}{}_{ij}.\end{aligned} \tag{39.14}$$

But

$$2\pi M\left(\frac{2m^*}{h^2}\right)^{3/2}(E-E_0)^{1/2}=N(E) \tag{39.15}$$

is the density of states in energy, therefore we represent (39.14) in the form

$$\int\limits_{(S_E)}\frac{v_iv_j}{\hbar v}dS_E=\frac{2}{3m^2}(E-E_0)\,8\pi^3N(E)\,\delta_{ij}. \tag{39.16}$$

Substituting (39.16) into (39.15) we obtain (for $E_0=0$)

$$K'^{ij}_{rs}=-\frac{4\delta_{ij}}{3m^*}\int\limits_0^{\infty}\frac{E^r\tau^s}{1+\mu^2B^2}N(E)\frac{\partial f_0}{\partial E}dE. \tag{39.17}$$

It follows from expression (39.17) that *all the diagonal elements $K_{rs}^{\prime ij}$ are identical*; this means that *in the case of spherical constant-energy surfaces the kinetic coefficients are scalars*: $K_{rs}^{\prime ij}=K'_{rs}$.

1. Non-degenerate semiconductors. In a non-degenerate semiconductor the Fermi-Dirac function reduces to the Boltzmann function, for which

$$\frac{\partial f_0}{\partial E}=-\frac{f_0}{kT} \qquad (39.18)$$

and

$$K'_{rs}=\frac{4}{3m^*}\int\limits_0^\infty \frac{E^r\tau^s}{1+\mu^2B^2}f_0(E, T)\,N(E)\frac{dE}{kT}. \qquad (39.19)$$

Let $r=s=0$, and $B=0$. The integral for K_{00} is easily calculated:

$$K_{00}=\frac{4}{3m^*}\int\limits_0^\infty f_0(E, T)\,N(E)\frac{dE}{kT}=\frac{2}{3m^*kT}n, \qquad (39.20)$$

where n is the charge carrier (electron or hole) concentration;

$$n=2\int\limits_0^\infty f_0(E, T)\,N(E)\,dE, \qquad (39.21)$$

i.e. K_{00} is expressed in terms of carrier concentration. But (39.19) shows K'_{rs} to be proportional to charge carrier concentration since $2f_0(E, T)\,N(E)\,dE$ is the number of particles in the energy interval dE, so that

$$K'_{rs}=\frac{2}{3m^*kT}\int\limits_0^\infty \frac{E^r\tau^s}{1+\mu^2B^2}dn(E). \qquad (39.22)$$

Multiply and divide (39.19) by the carrier concentration:

$$K'_{rs}=\frac{n}{m^*}\frac{2}{3kT}\frac{\int\limits_0^\infty \frac{E^r\tau^s}{1+\mu^2B^2}dn(E)}{\int\limits_0^\infty dn(E)}. \qquad (39.23)$$

Denote the multiplier at $\frac{n}{m^*}$ in (39.23) by

$$\frac{2}{3kT}\frac{\int\limits_0^\infty \frac{E^r\tau^s}{1+\mu^2B^2}dn(E)}{\int\limits_0^\infty dn(E)}=\left\langle\frac{E^r\tau^s}{1+\mu^2B^2}\right\rangle. \qquad (39.24)$$

This is the relaxation time averaged over the carriers with the weight $\frac{E^r}{1+\mu^2B^2}$ *and raised to the power of s.*

The expression (39.24) may be written in a form more convenient for calculation if use is made of a dimensionless variable $x=\frac{E}{kT}$:

$$\left\langle\frac{E^r\tau^s}{1+\mu^2B^2}\right\rangle=\frac{2\,(kT)^{r+1/2}}{3\,(kT)^{3/2}}\,\frac{\int\limits_0^\infty \frac{\tau^s e^{-x}}{1+\mu^2B^2}x^{r+1/2}\,dx}{\int\limits_0^\infty e^{-x}x^{1/2}dx}=$$

$$=\frac{2(kT)^{r-1}}{3}\,\frac{\int\limits_0^\infty \frac{\tau^s}{1+\mu^2B^2}\,e^{-x}x^{r+1/2}\,dx}{\int\limits_0^\infty e^{-x}x^{1/2}\,dx}. \tag{39.25}$$

The s power of the relaxation time averaged with the weight $\frac{E^r}{1+\mu^2B^2}$ is also denoted by the mean value of τ^s with r brackets

$$\left\langle\frac{E^r\tau^s}{1+\mu^2B^2}\right\rangle=\underbrace{\left\langle\ldots\left\langle\frac{\tau^s}{1+\mu^2B^2}\right\rangle\ldots\right\rangle}_{r}. \tag{39.26}$$

Thus, the kinetic coefficients may be expressed in the form

$$K'_{rs}=\frac{n}{m^s}\left\langle\frac{E^r\tau^s}{1+\mu^2B^2}\right\rangle. \tag{39.27}$$

If the semiconductor contains charge carriers of different types, using (39.27) we may calculate the kinetic coefficients for each type.

Consider one specific but nevertheless practically very important case, when *the relaxation time is a power function of the energy*:

$$\tau(E)=\tau_0E^p=\frac{\tau_0}{(kT)^p}\,x^p=\tau'_0x^p, \tag{39.28}$$

where τ_0 and τ'_0 are constants.

Substitute (39.28) into (39.25):

$$\left\langle\frac{E^r\tau^s}{1+\mu^2B^2}\right\rangle=\frac{2}{3}(kT)^{r-1}\tau_0'^s\,\frac{\int\limits_0^\infty \frac{x^{sp+r+1/2}}{1+\mu^2B^2}e^{-x}dx}{\int\limits_0^\infty x^{1/2}e^{-x}dx}. \tag{39.29}$$

Putting $B=0$ we reduce the integrals (39.29) to the Euler gamma-functions

$$\langle E^r\tau^s\rangle = (kT)^{r-1}\tau_0'^s\,\frac{\Gamma\left(sp+r+\frac{3}{2}\right)}{\Gamma\left(\frac{5}{2}\right)} \qquad (39.30)$$

The appearance of $\Gamma(^5/_2)$ instead of $\Gamma(^3/_2)$ in the denominator is due to the fact that instead of writing $^3/_2$ in the numerator we write $^3/_2$ in the denominator, and that $^3/_2\,\Gamma(^3/_2)=\Gamma(^5/_2)$. The expression (39.30) will be made use of below.

2. **Degenerate semiconductor.** For a degenerate semiconductor $-\frac{\partial f_0}{\partial E}\cong\delta(E-F)$, and this makes the calculation of the integral (39.17) quite easy:

$$K'_{rs}=\frac{4}{3m^*}\,\frac{F^r\tau^s(F)}{1+\mu^2(F)B^2}\,N(F)=$$
$$=\frac{8\pi}{3m^*}\left(\frac{2m^*}{h^2}\right)^{3/2}F^{3/2}\,\frac{F^{r-1}\tau^s(F)}{1+\mu^2(F)B^2}=\frac{n}{m^*}\,\frac{F^{r-1}\tau^s}{1+\mu^2(F)B^2}. \qquad (39.31)$$

Should we write K'_{rs} for degenerate semiconductors in the form (39.27), we would be able to define the mean relaxation time by the relation

$$\left\langle\frac{E^r\tau^s}{1+\mu^2B^2}\right\rangle=\frac{F^{r-1}\tau^s(F)}{1+\mu^2(F)B^2}, \qquad (39.32)$$

i.e. *the averaged relaxation time is determined solely by the relaxation time of charge carriers lying on the Fermi surface.*

Effective mass—rank II tensor. Now we are going to calculate the kinetic coefficients for the case of a tensor effective mass which, for the sake of simplicity, is presumed to be of a diagonal form:

$$K'^{ij}_{rs}=-\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})}\frac{E^{r-1}\tau^s v_i v_j}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\mathbf{m}^*\mathbf{B})}\,\frac{\partial f_0}{\partial E}\,d\tau_{\mathbf{k}}. \qquad (39.33)$$

We shall, as before, assume τ to be dependent only on the energy: $\tau=\tau(E)$; but now τ is going to be *anisotropic.* Express the element of volume $d\tau_{\mathbf{k}}$ in terms of the energy element dE and the element of the energy surface:

$$d\tau_{\mathbf{k}}=dS_E\,dk_n=\frac{dS_E\,dE}{\hbar\,|\mathbf{v}|}=\frac{dS_E\,dE}{\hbar\sqrt{\sum\limits_{l=1}^{3}\frac{\hbar^2k_l^2}{m_l^2}}}. \qquad (39.34)$$

To calculate the integral over the surface introduce different scales for different directions in the Brillouin zone, i.e. introduce new variables

$$w_l = \frac{\hbar k_l}{\sqrt{2m_l}}; \quad \hbar k_l = \sqrt{2m_l}\, w_l, \tag{39.35}$$

or

$$k_l = \frac{\sqrt{2m_l}}{\hbar} w_l. \tag{39.36}$$

Find the relation between the energy and w_l:

$$E = E_0 + \sum_{l=1}^{3} \frac{\hbar^2 k_l^2}{2m_l} = E_0 + \sum_{l=1}^{3} w_l^2 \tag{39.37}$$

It follows from (39.37) that in the w_l variables the constant-energy surface is a sphere.

Express K'_{rs} in $\mathbf{w}$. To this end write

$$d\tau_{\mathbf{k}} = dk_x dk_y dk_z = \frac{(8m_1m_2m_3)^{1/2}}{\hbar^3} d\tau_{\mathbf{w}}. \tag{39.38}$$

According to (39.5)

$$d\tau_{\mathbf{w}} = dS(\mathbf{w})\, dw_{\mathrm{n}}. \tag{39.39}$$

On the other hand,

$$dE = \left(\frac{dE}{d\mathbf{w}} d\mathbf{w}\right) = (\mathbf{w}\, d\mathbf{w}) = w\, dw_{\mathrm{n}} = \sqrt{E}\, dw_{\mathrm{n}}, \tag{39.40}$$

therefore

$$d\tau_{\mathbf{w}} = dS(\mathbf{w}) \frac{dE}{\sqrt{E}}, \tag{39.41}$$

or

$$d\tau_{\mathbf{k}} = \frac{(8m_1m_2m_3)^{1/2}}{\hbar^3} \frac{dE\, dS(\mathbf{w})}{\sqrt{E}}. \tag{39.42}$$

Taking into account that

$$v_l = \frac{\hbar k_l}{m_l} = \sqrt{\frac{2}{m_l}} w_l, \tag{39.43}$$

we write

$$K'^{ij}_{rs} = -\frac{1}{4\pi^3} \int_{(E)} \frac{\partial f_0}{\partial E} \frac{E^{r-1}\tau^s}{1 + \frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B}, \mathbf{m}^*\mathbf{B})} \frac{dE}{\sqrt{E}} \int_{(S)} \frac{(8\,|\mathbf{m}^*|)^{1/2}}{\hbar^3} \times$$
$$\times \sqrt{\frac{4}{m_i m_j}} w_i w_j dS = -\frac{2\,(8\,|\mathbf{m}^*|)^{1/2} M}{\hbar^3 \sqrt{m_i m_j}} \times$$
$$\times \int_{(E)} \frac{\partial f_0}{\partial E} \frac{E^{r-1}\tau^s w^4}{1 + \frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B}, \mathbf{m}^*\mathbf{B})} \frac{dE}{\sqrt{E}} \int_{(S_1)} \frac{w_i w_j}{w^2} d\Omega. \tag{39.44}$$

But according to (39.13)

$$\int\limits_{(S_1)} \frac{w_i w_j}{w^2}\, d\Omega = \frac{4\pi}{3}\,\delta_{ij}, \tag{39.45}$$

and, hence, K'_{rs} *is a diagonal tensor*. Introducing the expression for the density of states

$$N(E) = \frac{M2\pi}{h^3}(8m_1m_2m_3)^{1/2}E^{1/2}\ (E_0 = 0), \tag{39.46}$$

we may represent the kinetic coefficient in the form

$$K'^{ij}_{rs} = -\frac{4\delta_{ij}}{3\sqrt{m_i m_j}}\int\limits_0^\infty \frac{\partial f_0}{\partial E}\,\frac{N(E)\,E^r\tau^s}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\,\mathbf{m}^*\mathbf{B})}\,dE. \tag{39.47}$$

For an isotropic (scalar) effective mass the expression (39.47) reduces to (39.17).

The expression (39.47) for K'^{ij}_{rs} may be written in the following form:

$$K'^{ij}_{rs} = \frac{n\delta_{ij}}{m_i}\left\langle\frac{E^r\tau^s}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\,\mathbf{m}^*\mathbf{B})}\right\rangle, \tag{39.48}$$

since for $i = j$ $m_i = m_j$. It follows that *all the kinetic coefficients are anisotropic, and the tensor* K'_{rs} *is proportional to the tensor* $\mathbf{m}^{*-1}$:

$$K'_{rs} = n\left\langle\frac{E^r\tau^s}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\,\mathbf{m}^*\mathbf{B})}\right\rangle \mathbf{m}^{*-1}. \tag{39.49}$$

This means that the *shape of constant-energy surfaces appreciably affects the course of kinetic phenomena.*

Summary of Sec. 39

1. The calculation of kinetic coefficients may conveniently be reduced to integration over the energy surface with subsequent integration with respect to energy. The calculations are easier for the case of spherical constant-energy surfaces, therefore to simplify calculations the ellipsoidal energy surfaces are, by a suitable choice of scales along the ellipsoids axes, transformed into spherical surfaces.

2. The kinetic coefficient tensor is proportional to the reciprocal effective mass tensor.

3. The kinetic coefficient tensor is

$$K'_{rs}=n\mathbf{m}^{*-1}\left\langle \frac{E^r\tau^s}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\mathbf{m}^*\mathbf{B})}\right\rangle. \tag{39.1s}$$

In the absence of a magnetic field one may write, omitting the prime,

$$K_{rs}=n\mathbf{m}^{*-1}\langle E^r\tau^s\rangle=\frac{n\langle\ldots\langle\tau^s\rangle_{(r)}\ldots\rangle}{\mathbf{m}^*} \tag{39.2s}$$

4. The averaged relaxation time for non-degenerate semiconductors is determined by the expression

$$\left\langle \frac{E^r\tau^s}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\mathbf{m}^*\mathbf{B})}\right\rangle=$$

$$=(kT)^{r-1}\frac{\int\limits_0^\infty \frac{\tau^s(x)\,x^{r+1/2}}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\mathbf{m}^*\mathbf{B})}e^{-x}\,dx}{\int\limits_0^\infty x^{3/2}e^{-x}\,dx}. \tag{39.3s}$$

and for degenerate semiconductors the expression is

$$\left\langle \frac{E^r\tau^s}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\mathbf{m}^*\mathbf{B})}\right\rangle=\frac{F^{r-1}\tau^s(F)}{1+\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B},\mathbf{m}^*\mathbf{B})}. \tag{39.4s}$$

40. CONDUCTIVITY OF SEMICONDUCTORS

In order to describe the electric conductivity of semiconductors the relations should be established between the current $\mathbf{j}$ and the electric field $\mathbf{E}$ responsible for it. To this end one should put $\mathbf{B}=0$ and $\nabla F=\nabla T=0$. Then we may write, using (38.19),

$$\mathbf{j}=e^2K_{11}\mathbf{E}=\sigma\mathbf{E}. \tag{40.1}$$

It follows from (40.1) that *the relation between the conductivity tensor σ and the tensor K_{11} is*

$$e^2K_{11}=\sigma, \tag{40.2}$$

or

$$\sigma=e^2\frac{n\langle E\tau\rangle}{\mathbf{m}^*}=en\frac{e\langle\tau\rangle}{\mathbf{m}^*}=en\mu_d. \tag{40.3}$$

Thus, *the Boltzmann kinetic equation not only leads to Ohm's law, but elucidates the more profound meaning of conductivity, as well. Specific conductance and mobility are determined by the rela-*

xation time averaged with the weight E:

$$\langle\tau\rangle=\langle E\tau\rangle=\frac{\int\limits_0^\infty x^{3/2}e^{-x}\tau(x)\,dx}{\int\limits_0^\infty x^{3/2}e^{-x}\,dx}. \qquad (40.4)$$

When relaxation time depends on some power of energy $\tau=\tau_0E^p$, or $\tau=\tau_0'x^p$, (40.4) enables us to write *for a non-degenerate semiconductor*:

$$\langle\tau\rangle=\tau_0'\frac{\Gamma\left(\frac{5}{2}+p\right)}{\Gamma\left(\frac{5}{2}\right)}. \qquad (40.5)$$

Table 9 shows the ratio $\frac{\Gamma\left(\frac{5}{2}+p\right)}{\Gamma\left(\frac{5}{2}\right)}$ for some values of p.

Table 9

p	0	1/2	1	3/2	2	5/2	−1/2	−1	−3/2	−2
$\Gamma\left(\frac{5}{2}+p\right)$	$\frac{3\sqrt{\pi}}{4}$	2	$\frac{15\sqrt{\pi}}{8}$	6	$\frac{105\sqrt{\pi}}{16}$	24	1	$\frac{\sqrt{\pi}}{2}$	1	$\sqrt{\pi}$
$\frac{\Gamma\left(\frac{5}{2}+p\right)}{\Gamma\left(\frac{5}{2}\right)}$	1	$\frac{8}{3\sqrt{\pi}}$	32/5	$8/\sqrt{\pi}$	35/4	$\frac{32}{\sqrt{\pi}}$	$\frac{4}{3\sqrt{\pi}}$	2/3	$\frac{4}{3\sqrt{\pi}}$	4/3

For degenerate semiconductors

$$\langle\tau\rangle=\tau(F), \qquad (40.6)$$

therefore drift mobility is determined only by relaxation time of electrons or holes occupying states near the Fermi surface, the conductivity being dependent on electron concentration. This paradox may be readily resolved if the expression for the distribution function is considered:

$$f=f_0+f^{(1)}=f_0-\frac{\partial f_0}{\partial E}(\mathbf{X}\mathbf{v})=f_0-\frac{\partial f_0}{\partial E}e(\mathbf{E}\mathbf{v})\tau. \qquad (40.7)$$

For degenerate semiconductors $\frac{\partial f_0}{\partial E}\cong0$ for $E\neq F$, and the correction to the unperturbed function f_0 is, therefore, zero. In other words, *the external field* **E** *perturbs only the states in a strip within* $\pm kT$ *of the Fermi surface.* Since relaxation should take place only for those states, the mean relaxation time $\langle\tau\rangle$ should be determined by the value of $\tau(F)$.

For non-degenerate semiconductors

$$-\frac{\partial f_0}{\partial E}=\frac{f_0}{kT} \tag{40.8}$$

and

$$f=f_0+f^{(1)}=f_0+\frac{f_0}{kT}e(\mathbf{Ev})\tau, \tag{40.9}$$

i.e.

$$f^{(1)}=f_0\frac{e}{kT}(\mathbf{Ev})\tau. \tag{40.10}$$

The perturbation is different for different states: the lower is the energy, the greater is f_0 and the greater should $f^{(1)}$ be. However, since for $E=E_c$ $\mathbf{v}=0$, the perturbation of the states lying close to E_c will be less than that of the states lying somewhat higher (at a distance of the order of kT above E_c). Consider the relative change in the distribution function

$$\frac{f-f_0}{f_0}=\frac{f^{(1)}}{f_0}=\frac{e}{kT}(\mathbf{Ev})\tau. \tag{40.11}$$

We see that, as energy E increases, so does the relative change in the distribution function, and with it $\mathbf{v}$. Moreover, the relative change $\frac{f^{(1)}}{f_0}$ will be the higher, the lower will the temperature be. The meaning of this is obvious. The change in f results from the work dA performed by the force $e\mathbf{E}$ acting on the electrons moving in the electric field $\mathbf{E}$. In the spell of time dt electron covers the distance $ds=\mathbf{v}\,dt$.

The potential difference between two points at a distance ds from each other is

$$d\varphi=-(\mathbf{E}ds)=-(\mathbf{Ev})\,dt, \tag{40.12}$$

and the potential energy difference

$$dV=ed\varphi=-e(\mathbf{Ev})\,dt. \tag{40.13}$$

If one takes into account that the total energy of the field-electron system remains constant, then

$$dE+dV=0, \tag{40.14}$$

$$dE=-dV=e(\mathbf{Ev})\,dt=dA;\quad dA=-dV, \tag{40.15}$$

i.e. the energy of an electron moving along the field $(\mathbf{Ev}) > 0$ decreases (because $e < 0$) and it goes over to lower energy levels. The energy of an electron moving against the field increases as $(\mathbf{Ev}) < 0$ and $dE > 0$, and it goes over to higher energy levels. Thus, the electric field increases the energy of the electrons which possess a velocity component directed against the field, and decreases the energy of the electrons which possess a velocity component directed along the field. In compliance with (40.11) this leads to an increase in the number of particles with speeds directed against the field and to a decrease in the number of particles with speeds (or speed components) directed along the field.

The change in the number of particles will be the greater the smaller is f, i.e. the smaller is the occupancy of the energy levels. The occupation of the levels makes it impossible for electrons to move from level to level, therefore for $E < F$ in degenerate semiconductors $f^{(1)} = 0$. This, however, means that *only the electrons of a narrow energy strip of the order of* $F \pm kT$ *take part in the current in degenerate semiconductors (and metals)*.

The number of electrons close to the Fermi surface will be the greater the greater is the surface area S_F. S_F, in turn, will be greater for higher electron concentrations n. As a result, the number of effective charge carriers will be proportional to the total number of electrons n, and this explains the contradiction between the above considerations and the formula

$$\sigma = en\mu(F) = e^2\tau(F)\, n\mathbf{m}^{*-1} \tag{40.16}$$

which includes the total number of electrons.

Consider now the conductivity of an M-valley (to be precise, M-energy ellipsoid) semiconductor.

Denoting the valley number by ν and the current density of corresponding charge carriers by $\mathbf{j}^{(\nu)}$ we obtain the total current density

$$\mathbf{j} = \sum_{\nu=1}^{M} \mathbf{j}^{(\nu)}. \tag{40.17}$$

To find $\mathbf{j}^{(\nu)}$ the vector $\mathbf{E}$ should be represented in the main axes of the $\mathbf{m}^{*-1}$ tensor for the νth valley. Suppose $\mathbf{E}$ is of the form

$$\mathbf{E} = (E_1^{(\nu)};\ E_2^{(\nu)};\ E_3^{(\nu)}). \tag{40.18}$$

In this case

$$\mathbf{j}^{(\nu)} = (j_1^{(\nu)};\ j_2^{(\nu)};\ j_3^{(\nu)}) = (\sigma_1^{(\nu)}E_1^{(\nu)};\ \sigma_2^{(\nu)}E_2^{(\nu)};\ \sigma_3^{(\nu)}E_3^{(\nu)}), \tag{40.19}$$

where the form of

$$\sigma_i^{(\nu)} = \frac{e^2n^{(\nu)}}{m_i}\langle\tau\rangle \tag{40.20}$$

is similar for all valleys. In our notation $n^{(\nu)}$ denotes the number of charge carriers in the νth valley (in the νth ellipsoid). Since all the ellipsoids are equivalent, it follows from the condition

$$\sum_{\nu=1}^{M} n^{(\nu)} = n \tag{40.21}$$

that

$$n^{(\nu)} = \frac{n}{M}. \tag{40.22}$$

By transforming all the conductivity tensors $\sigma^{(\nu)}$ to the same co-ordinate axes one may reduce the current density $\mathbf{j}$ to the form

$$\mathbf{j} = \sum_{\nu=1}^{M} \sigma^{(\nu)} \mathbf{E} = \sigma \mathbf{E}, \tag{40.23}$$

where the conductivity tensor σ is the sum of the $\sigma^{(\nu)}$ tensors. As is well known, the ijth component of a tensor sum is equal to the sum of the ij components of the addends which should be written

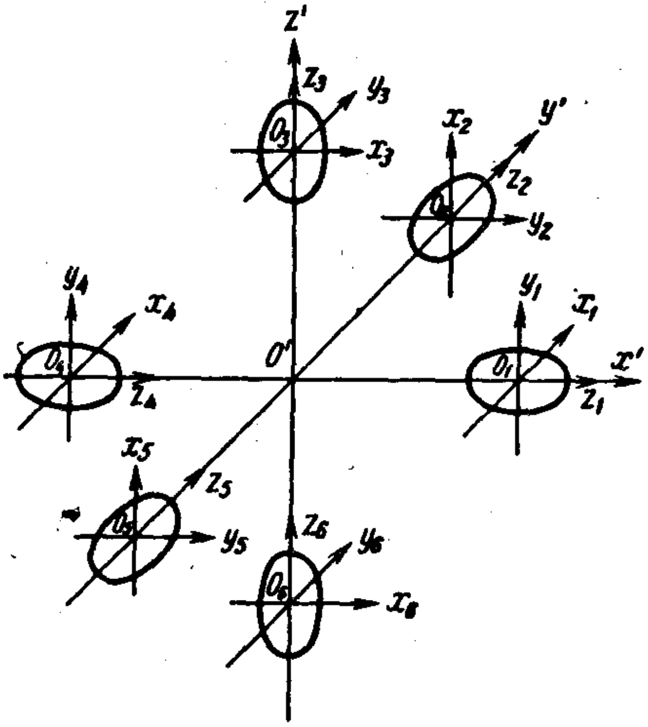


Fig. 60. The relation between the co-ordinate axes of various constant-energy surfaces of the conduction band in silicon

in the same co-ordinate system. As a reminder: the expression for a tensor **T** given in one co-ordinate system $\{x_i\}$ in another coordinate system $\{x'_i\}$ is

$$T'_{ij} = \sum_{lm} \frac{\partial x'_i}{\partial x_l} \frac{\partial x'_j}{\partial x_m} T_{lm}. \tag{40.24}$$

If

$$x_i' = \sum_l a_{il} x_l, \tag{40.25}$$

it follows that

$$\frac{\partial x_i'}{\partial x_l} = a_{il} = \cos(x_i', x_l) \tag{40.26}$$

and

$$T'_{ij} = \sum_{lm} a_{il} a_{jm} T_{lm}. \tag{40.27}$$

The expression for the tensor $\mathbf{T}'$ in the $\{x'\}$ co-ordinate system may be found if the position of both co-ordinate systems is known. Consider the more simple case of silicon as an example.

We choose the three [100] directions as co-ordinate axes (x', y', z') *of the crystal as a whole* (Fig. 60). For the ellipsoids *1* and *4* we choose for the main (x, y, z) axes:

$$\begin{aligned} x' &= z^{(1)} = z^{(4)}, \\ y' &= x^{(1)} = x^{(4)}, \\ z' &= y^{(1)} = y^{(4)}. \end{aligned} \tag{40.28}$$

The axes transformation matrix may be written in the form

$$\mathbf{A}^{(1)} = \mathbf{A}^{(4)} = \{a_{ij}^{(1)}\} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}. \tag{40.29}$$

For the ellipsoids *2* and *5* we may write the following axes transformations:

$$x' = y^{(2)} = y^{(5)}; \quad y' = z^{(2)} = z^{(5)}; \quad z' = x^{(2)} = x^{(5)};$$

$$\mathbf{A}^{(2)} = \mathbf{A}^{(5)} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \tilde{\mathbf{A}}^{(1)}. \tag{40.30}$$

For the ellipsoids *3* and *6* the axes with and without the strokes coincide:

$$x' = x^{(3)} = x^{(6)}; \quad y' = y^{(3)} = y^{(6)}; \quad z' = z^{(3)} = z^{(6)};$$

$$\mathbf{A}^{(3)} = \mathbf{A}^{(6)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{40.31}$$

The choice of the axes transformation matrix is dictated by the following obvious considerations: the third main axis corresponds to the ellipsoid rotation axis, therefore

$$z^{(1)} = z^{(4)} = x'; \quad z^{(2)} = z^{(5)} = y'; \quad z^{(3)} = z^{(6)} = z'. \tag{40.32}$$

The position of the first and the second axes in the plane perpendicular to the rotation axis is arbitrary. To simplify the transformations they may be conveniently made to coincide with the stroked axes, so that only the right-hand screw coordinate system is retained. Write the tensors $\sigma^{(\nu)}$ in the (x', y', z') axes using the rule of transformation (40.27):

$$\sigma_{ij}'^{(\nu)} = \sum_{lm} a_{il}a_{jm}\sigma_{lm}^{(\nu)} = \sum_{lm} a_{il}a_{jm}\sigma_l^{(\nu)}\delta_{lm} = \sum_l a_{il}a_{jl}\sigma_l^{(\nu)}. \quad (40.33)$$

Leaving out detailed calculations we write the tensors $\sigma'^{(\nu)}$ [taking into account (40.33) and (40.29-31)]:

$$\sigma'^{(1)} = \sigma'^{(4)} = \begin{pmatrix} \sigma_3^{(1)} & 0 & 0 \\ 0 & \sigma_1^{(1)} & 0 \\ 0 & 0 & \sigma_2^{(1)} \end{pmatrix};$$

$$\sigma'^{(2)} = \sigma'^{(5)} = \begin{pmatrix} \sigma_2^{(2)} & 0 & 0 \\ 0 & \sigma_3^{(2)} & 0 \\ 0 & 0 & \sigma_1^{(2)} \end{pmatrix};$$

$$\sigma'^{(3)} = \sigma'^{(6)} = \begin{pmatrix} \sigma_1^{(3)} & 0 & 0 \\ 0 & \sigma_2^{(3)} & 0 \\ 0 & 0 & \sigma_3^{(3)} \end{pmatrix}. \quad (40.34)$$

The form of the tensors $\sigma'^{(\nu)}$ may be easily obtained if it is taken into account that the product of the elements of any two different $(i \neq j)$ lines of the same column (l) for all the three matrices $\mathbf{A}$ is zero, and, therefore, σ' remain diagonal. Simple physical considerations will help us understand why, for example, the form of $\sigma'^{(1)}$ is as shown in (40.34). If electric field is applied along the x' axis it will coincide with the direction of the rotational axes of the first and the fourth ellipsoids, and the current should, therefore, be determined by the σ_3 component. For the other two pairs of ellipsoids the electric field lies in the plane perpendicular to the rotation axis, therefore, their contribution to conductivity will be determined by the components σ_1 and σ_2. Find the tensor of full conductivity which will also be of a diagonal form:

$$\sigma_{ij} = \sigma_i\delta_{ij},$$

where

$$\begin{aligned} \sigma_1 &= 2\,(\sigma_1^{(3)} + \sigma_2^{(2)} + \sigma_3^{(1)}), \\ \sigma_2 &= 2\,(\sigma_1^{(1)} + \sigma_2^{(3)} + \sigma_3^{(2)}), \\ \sigma_3 &= 2\,(\sigma_1^{(2)} + \sigma_2^{(1)} + \sigma_3^{(3)}). \end{aligned} \quad (40.35)$$

We see from here, that the diagonal components of full conductivity are obtained by summing up the diagonal elements of the different conductivity ellipsoids. If these ellipsoids are equivalent all $\sigma^{(v)}$ will be identical, and

$$\begin{aligned}\sigma_1 &= 2\,(\sigma_1^{(v)}+\sigma_2^{(v)}+\sigma_3^{(v)}) = \sigma,\\ \sigma_2 &= 2\,(\sigma_1^{(v)}+\sigma_2^{(v)}+\sigma_3^{(v)}) = \sigma,\\ \sigma_3 &= 2\,(\sigma_1^{(v)}+\sigma_2^{(v)}+\sigma_3^{(v)}) = \sigma.\end{aligned} \tag{40.36}$$

The specific electric conductance is described by a diagonal tensor with equal components, i.e. σ is a scalar. Hence, *a symmetrical distribution of anisotropic valley conductivities results in an isotropic full conductivity*. Express the conductivity σ in terms of carrier concentration and mobility:

$$\begin{aligned}\sigma = 2\,(\sigma_1^{(v)}+\sigma_2^{(v)}+\sigma_3^{(v)}) &= 2e^2n^{(v)}\langle\tau\rangle\left(\frac{1}{m_1}+\frac{1}{m_2}+\frac{1}{m_3}\right) =\\ &= 2e^2n\,\langle\tau\rangle\left(\frac{2}{m_t}+\frac{1}{m_l}\right).\end{aligned} \tag{40.37}$$

We took into account that $m_1=m_2=m_t$ and $m_3=m_l$. Introduce *an "isotropic" effective mass for conductivity* $\tilde{m}^*$ using the relation

$$\frac{1}{\tilde{m}^*}=\frac{1}{3}\left(\frac{1}{m_1}+\frac{1}{m_2}+\frac{1}{m_3}\right)=\frac{1}{3}\left(\frac{2}{m_t}+\frac{1}{m_l}\right). \tag{40.38}$$

With $\tilde{m}^*$ we may write the expression for σ in the form

$$\sigma=\frac{6e^2n^{(v)}\langle\tau\rangle}{\tilde{m}^*}=\frac{ene\,\langle\tau\rangle}{\tilde{m}^*}=en\mu_d, \tag{40.39}$$

where

$$\mu_d=\frac{e\,\langle\tau\rangle}{\tilde{m}^*}=\frac{e\,\langle\tau\rangle}{3}\left(\frac{2}{m_t}+\frac{1}{m_l}\right). \tag{40.40}$$

We see from here that *the mobility is determined by the "isotropic" effective mass $\tilde{m}^*$ which is, itself, related to the shape of constant-energy surfaces.* $\tilde{m}^{*-1}$ is the arithmetic mean of m_i^{-1}, while m_d^{-1} is related to the geometric mean of the components m_i^{-1}:

$$m_d^3=M^2 8m_1m_2m_3=M^2 8m_t^2m_l. \tag{40.41}$$

The conductivity of n-type germanium may be considered in a similar fashion.

Consider the conductivity of a substance containing charge carriers of different kinds, i.e. electrons and holes of different effective mass. Denoting the current density due to the charge carriers of the α-kind by $\mathbf{j}_\alpha$ we obtain the total current density $\mathbf{j}$ in the form

$$\mathbf{j}=\sum_\alpha \mathbf{j}_\alpha=\sigma\mathbf{E}=\left(\sum_\alpha\sigma_\alpha\right)\mathbf{E}, \tag{40.42}$$

or

$$\sigma=\sum_{\alpha}\sigma_{\alpha}=\sum_{\alpha}e_{\alpha}^{2}K_{11(\alpha)}=e^{2}\sum_{\alpha}\frac{n_{\alpha}\langle\tau_{\alpha}\rangle}{\overset{\bullet}{\mathrm{m}}_{\alpha}}=\sum_{\alpha}e_{\alpha}n_{\alpha}\mu_{d\alpha}. \quad (40.43)$$

Let us now turn to the equation (38.2s) for energy flux density. In the assumptions made in the derivation of the expression (40.1) we obtain

$$\mathbf{W}=eK_{21}\mathbf{E}. \quad (40.44)$$

Express the energy flux density **W** in terms of the current density **j**:

$$\mathbf{W}=eK_{21}\mathbf{E}=eK_{21}\frac{\mathbf{j}}{e^{2}K_{11}}=\Pi\mathbf{j}. \quad (40.45)$$

The energy flux caused by the directional motion of charge carriers is termed Peltier current; Π being the Peltier coefficient. It follows from (40.45) that the Peltier coefficient is expressed in terms of the kinetic coefficients:

$$\Pi=\frac{K_{21}}{eK_{11}}=\frac{1}{e}\frac{\langle E^{2}\tau\rangle}{\langle E\tau\rangle}. \quad (40.46)$$

Summary of Sec. 40

1. The expression for current density in a uniform semiconductor all the points of which are at the same temperature in the absence of magnetic field is

$$\mathbf{j}=e^{2}K_{11}\mathbf{E}, \quad (40.1s)$$

i.e. the kinetic coefficient K_{11} is equal to the specific conductance tensor σ divided by the square of the charge.

2. The averaged relaxation time $\langle\tau\rangle$ in a non-degenerate semiconductor for the case $\tau=\tau_{0}E^{p}=\tau_{0}'x^{p}$ is expressed with the aid of the Γ-function:

$$\langle\tau\rangle=\tau_{0}'\frac{\Gamma\left(\frac{5}{2}+p\right)}{\Gamma\left(\frac{5}{2}\right)}. \quad (40.2s)$$

3. In a degenerate semiconductor the averaged relaxation time coincides with the relaxation time of charge carriers occupying the Fermi surface:

$$\langle\tau\rangle\cong\tau(F). \quad (40.3s)$$

4. It follows from the kinetic equation that the drift mobility of charge carriers is determined by the relaxation time averaged

over energy:

$$\mu_d = \frac{e\langle\tau\rangle}{m^*}. \qquad (40.4s)$$

5. If the semiconductor has M equivalent valleys its full conductivity σ will be equal to the sum of the conductivities of all the valleys:

$$\sigma = \sum_{\nu=1}^{M} \sigma^{(\nu)}. \qquad (40.5s)$$

The tensors $\sigma^{(\nu)}$ should in this case be written in the same coordinate system. If the conductivity of each valley is anisotropic the full conductivity σ may, however, be isotropic provided the valleys are symmetrically arranged in the Brillouin zone. Full conductivity may be described with the aid of the conductivity effective mass $\tilde{m}^*$ which for the conduction band in silicon is related to the effective mass tensor components by the expression

$$\frac{1}{\tilde{m}^*} = \frac{1}{3}\left(\frac{1}{m_1} + \frac{1}{m_2} + \frac{1}{m_3}\right) = \frac{1}{3}\left(\frac{2}{m_t} + \frac{1}{m_l}\right). \qquad (40.6s)$$

41. GALVANOMAGNETIC EFFECTS

The physical phenomena which take place in a substance placed in a magnetic field in the presence of electric current due to the action of an electric field are termed galvanomagnetic. In other words, *galvanomagnetic phenomena are the result of combined action of the electric and the magnetic field.* Galvanomagnetic phenomena include: (1) the Hall effect; (2) the magnetoresistive effect; (3) the Ettingshausen effect, or the transverse galvanothermomagnetic effect; (4) the Nernst effect, or the longitudinal galvanothermomagnetic effect. The Hall effect in the narrow sense of this word is also termed galvanomagnetic. The terms cited above—the 'transverse' and the 'longitudinal' galvanothermomagnetic effect—reflect the direction of the temperature gradient relative to the direction of the current. In respect to the direction of the magnetic field this gradient may be either transverse or longitudinal.

A qualitative description of the galvanomagnetic effect follows from the consideration of the motion of a charged particle in combined electric and magnetic fields under the action of the Lorentz force:

$$\mathbf{F} = e\mathbf{E} + e[\mathbf{vB}] = \mathbf{m}^*\ddot{\mathbf{r}}. \qquad (41.1)$$

Recall that in parallel electric and magnetic fields the particle moves in a helix with a continuously increasing pitch. This is quite understandable, since in a magnetic field alone the particle whose speed along the field is $v_{\parallel}$, and perpendicular to the field $v_{\perp}$, moves in a circle with the radius

$$r=\frac{m^{*}v_{\perp}}{eB}=\frac{v_{\perp}}{\omega_c}. \tag{41.2}$$

with the angular velocity $\omega_c=\frac{eB}{m^{*}}$ and travels along the field with a velocity $v_{\parallel}$.

In transverse (or crossed) **E** and **B** fields a particle of zero initial velocity moves along a cycloid: it rotates in a circle of radius

$$r=\frac{m^{*}E}{eB^{2}}, \tag{41.3}$$

the centre of which travels with a constant drift velocity

$$\mathbf{u}_d=\frac{[\mathbf{EB}]}{B^{2}} \tag{41.4}$$

in a direction perpendicular both to the electric and magnetic fields.

The path of a particle whose initial velocity $\mathbf{v}_0$ lies in a plane perpendicular to the magnetic field is a trochoid (an elongated or contracted cycloid). Both the electric and the magnetic fields do not affect the component of a particle velocity directed along the magnetic field, and this should be kept in mind when discussing the motion of a particle with such an initial velocity.

When dealing with particles moving in a solid one should take into account the collisions which disturb directional motion of particles initiated by the fields. *After each collision the particle will move in a helix or trochoid with new parameters.*

To assess the value of the field the relaxation time should be compared to the period of rotation of the particle in the magnetic field. If the relaxation time considerably exceeds the period $\frac{2\pi}{\omega_c}$, the particle will during the time τ make several revolutions while moving in a cycloid or in a helix. This happens in strong magnetic fields. Magnetic fields in which the particle during the time τ makes less than one revolution are considered to be weak. Thus, *in strong magnetic fields*

$$\frac{\omega_c\tau}{2\pi}=\frac{e\tau}{m^{*}}\frac{B}{2\pi}\gg 1. \tag{41.5}$$

In weak fields

$$\frac{\omega_c\tau}{2\pi}=\frac{e\tau}{m^{*}}\frac{B}{2\pi}\ll 1. \tag{41.6}$$

As may be seen from (41.5) and (41.6), *the definition of "strong" and "weak" fields depends not only on the magnetic field induction B, but on the charge carrier mobility, as well.* The conditions (41.5) and (41.6) may be related to the radius of the circumference r along which the particle is moving, and the mean free path l:

$$l = v\tau;\ r = \frac{v}{\omega_c};\quad \frac{l}{r} = \omega_c \tau. \tag{41.7}$$

Hence, *in weak magnetic fields $r \gg l$, i.e. the curvature of the particle path is negligible. In strong magnetic fields it is very great.*

To understand some phenomena it is sufficient to take into account just the drift velocity $\mathbf{v}_d = \mu_d \mathbf{E}$, others require the consideration of *the spread in the velocities of electrons.* All these peculiarities are reflected in the kinetic equation which by force of this provides the most accurate description of kinetic effects.

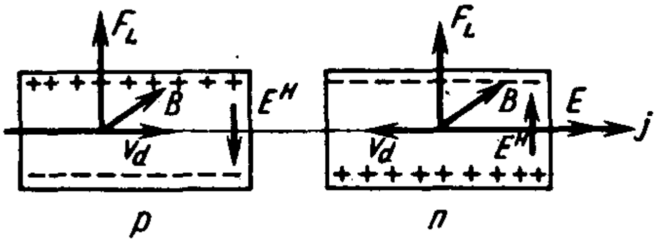


Fig. 61. The formation of the Hall field in an electron-type and a hole-type semiconductor

1. Hall effect. Codsider the qualitative aspects of the action of the magnetic field on electric current flowing in a semiconductor. Suppose the semiconductor sample is a parallelepiped of cross section $a \cdot c$ (Fig. 61). The electric field is directed along the x-axis: $\mathbf{E} = (E, 0, 0)$, the magnetic—along the y-axis: $\mathbf{B} = (0, B, 0)$. At the moment the electric field is switched on an electric current is established the density of which is

$$\mathbf{j} = \sigma \mathbf{E}. \tag{41.8}$$

The charge carriers acquire a directional velocity $\mathbf{v}_d$ (drift velocity) in the direction of the field in case of holes and against the field in case of electrons. When the magnetic field is switched on a force perpendicular both to $\mathbf{v}_d$ and $\mathbf{B}$ begins to act on the electrons and holes

$$\mathbf{F} = e[\mathbf{v}_d \mathbf{B}]. \tag{41.9}$$

But

$$\mathbf{v}_d = \mu_d \mathbf{E} = \frac{e\langle\tau\rangle}{m^*} \mathbf{E}, \tag{41.10}$$

therefore

$$\mathbf{F} = \frac{e^2 \langle\tau\rangle}{m^*} [\mathbf{EB}], \tag{41.11}$$

i.e. the *Lorentz force is independent of the charge carriers sign, being dependent only on the direction of* **E** *and* **B**, *or* **j** *and* **B**. In Fig. 61 **F** is directed upwards.

The electrons and holes whose velocity is determined by the electric field are deflected in the same direction.

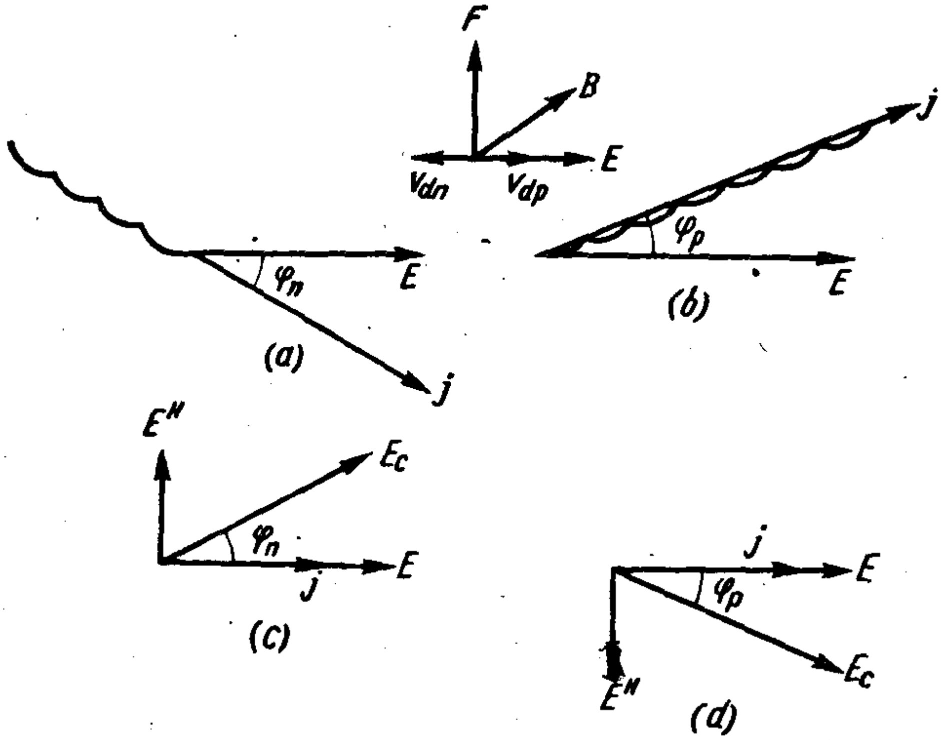


Fig. 62. The Hall angle in an infinite (*a*, *b*) and a finite (*c*, *d*) semiconductor: *a*, *c*—*n*-type; *b*, *d*—*p*-type

Thus, as a result of the action of the fields **E**, **B**, and of collisions, the electrons and holes will move in a straight line joining the ends of the cycloid sections and inclined by the angle φ to the field **E**. In other words, *the vector* **j** *will make an angle* φ *with the vector* **E**, *the direction of rotation being dependent on the charge carriers sign* (Fig. 62*a*, *b*) *just because both electrons and holes are deflected in the same direction*. This is the course the effect should take in an infinite semiconductor. If the dimensions of the semiconductor sample in the direction of the *z*-axis are finite, the component j_z at first will not be zero, and this will lead to charge carrier accumulation on the upper face (in our case) of the sample and to their deficit on the lower face. *The opposite faces of the sample will become charged, and as a result an electric field transverse to* **F** *will be established. This field is termed the Hall field, and this phenomenon is called the Hall effect. The direction of the Hall field* $\mathbf{E}^H$ *depends on the charge carriers sign*. In our example $\mathbf{E}^H$ is directed upwards in an *n*-type sample, and downwards in a *p*-type sample. Until the magnetic field has been applied, the equipotential planes in the sample were perpendicular to the *x*-axis, i.e. to the vector **j**. The value of $\mathbf{E}^H$ will continue to grow until the Lorentz force (41.9) is compensated by the transverse electric field. After this the charge carriers will move as if only the field **E** acted upon them, and their path will again become a straight line coinciding with the x-axis (recall that we are discussing only the directional motion). Thus, vector **j** will be directed along the field **E**. However, the resultant electric field $\mathbf{E}' = \mathbf{E} + \mathbf{E}^H$ will make an angle φ with the x-axis, or with **j** (Fig. 62*c*, *d*).

It follows, therefore, *that in an infinite semiconductor subject to rotation is the current vector, in a finite sample—the electric field vector. In both cases, however,* **j** *and* **E′** (or **E**) *make an angle* φ *termed Hall's angle.* Equipotential surfaces in a finite sample turn from their original position through the angle φ, therefore a potential difference $V^H = cE^H$ (where E^H is the Hall field intensity, c—the dimension of the sample in the direction perpendicular both to **E** and **B**) is established between extreme points lying in the same plane perpendicular to **j**. V^H *is termed Hall's potential difference.*

Hall found from experiment that $\mathbf{E}^H$ is determined by the current density **j** and the magnetic field induction **B**, as well as by the substance. *Corresponding properties of a substance are described by a quantity R termed Hall's coefficient. The four quantities* $\mathbf{E}^H$, **j**, **B** *and R are related through an empirical expression*

$$\mathbf{E}^H = R[\mathbf{B}\mathbf{j}] = -R[\mathbf{j}\mathbf{B}]. \qquad (41.12)$$

The quantity R may be easily found if it is taken into account that the Hall field should compensate the Lorentz force:

$$e\mathbf{E}^H + \mathbf{F} = 0. \qquad (41.13)$$

It follows from (41.13) that

$$\mathbf{E}^H = -\frac{1}{e}\mathbf{F} = -\frac{1}{e}e[\mathbf{v}_d\mathbf{B}] = -[\mathbf{v}_d\mathbf{B}] = -\mu_d[\mathbf{E}\mathbf{B}]. \qquad (41.14)$$

On the other hand, according to (41.12),

$$\mathbf{E}^H = -R[\mathbf{j}\mathbf{B}] = -R\sigma[\mathbf{E}\mathbf{B}]. \qquad (41.15)$$

Comparing (41.14) and (41.15) we see that

$$R\sigma = \mu_d \qquad (41.16)$$

and

$$R = \frac{\mu_d}{\sigma} = \frac{1}{en}. \qquad (41.17)$$

Here n was used to denote the charge carrier concentration (of electrons or holes). We see from here that *the Hall coefficient is inversely proportional to the charge carrier concentration*, and that *its sign coincides with the sign of the charge carriers*. Having established the sign of R we can find the sign of the charge carriers, or the conductivity type. The sign of R, in turn, is determined by the sign of $\mathbf{E}^H$, or V^H, if V^H is appropriately defined.

The Hall angle φ may be determined from the relation:

$$\tan\varphi = \frac{\mathbf{E}^H}{\mathbf{E}} = -\frac{RB\sigma\mathbf{E}}{\mathbf{E}} = -R\sigma B = -\mu_d B. \qquad (41.18)$$

The Hall field for given **E** and **B** depends only on the charge carriers mobility.

In a sample of mixed conductivity the Hall coefficient should depend on the properties of the charge carriers of both types. Evidently, if in an intrinsic semiconductor the mobilities of charge carriers of both types are the same, the Hall field will be zero.

The Hall coefficient should also depend on temperature. From the temperature dependence of the Hall coefficient the temperature dependence of the charge carrier concentration may be experimentally determined:

$$\ln n = -\ln(e_n R_n);\ \ln p = -\ln(e_p R_p). \qquad (41.19)$$

Assess the order of magnitude of R. Let $n = 10^{16}\,\text{cm}^{-3}$. Then

$$R = \frac{1}{1.6\times10^{-19}10^{16}}\,\frac{\text{cm}^3}{\text{C}} \cong 600\,\frac{\text{cm}^3}{\text{C}} = 6\times10^{-4}\,\frac{\text{m}^3}{\text{C}}.$$

Table 10 contains the values of R (in cm³/C) for some carrier concentrations (in cm⁻³).

Table 10

n, cm^{-3}	10^{10}	10^{13}	10^{15}	$6.25\times$ $\times10^{15}$	10^{16}	$6.25\times$ $\times10^{16}$	10^{17}	$6.25\times$ $\times10^{17}$	
R, cm^3·C^{-1}	$6.25\times$ $\times10^{8}$	$6.25\times$ $\times10^{5}$	$6.25\times$ $\times10^{3}$	10^{3}	625	100	62.5	10	
n, cm^{-3}	10^{18}	$6.25\times$ $\times10^{18}$	10^{19}	$6.25\times$ $\times10^{19}$	10^{20}	$6.25\times$ $\times10^{20}$	10^{21}	$6.25\times$ $\times10^{21}$	10^{22}
R, cm^3·C^{-1}	6.25	1	0.625	0.1	0.063	0.01	0.006	0.001	0.0006

We defined R by means of the field $\mathbf{E}^H$. However, the quantity measured in experiment is $V^H = E^H c$. The current $I = jac$ is measured instead of the current density. Therefore, R may be expressed in the form

$$R=\frac{E^H}{jB}=\frac{V^H}{c\frac{I}{ac}B}=\frac{V^Ha}{IB}. \tag{41.20}$$

If volt, metre, ampere, tesla are used to measure the quantities in the formula (41.20), R will be in m³/C. If units not belonging to any one system are used (volt, centimetre, ampere, gauss) and R is expressed in cm³/C, we obtain, taking into account that 1 m = 10² cm, 1 T = 10⁴ Gs, that

$$R\left(\frac{\text{cm}^3}{\text{C}}\right)=10^8\frac{V^H(V)\,a\,(\text{cm})}{I\,(\text{A})\,B\,(\text{Gs})}. \tag{41.21}$$

2. Magnetoresistance. Gauss effect. The application of a magnetic field results not only in the appearance of the Hall angle between **j** and **E**, but affects conductivity, as well. Without the magnetic field a particle moves in a straight line and travels a distance between two collisions equal to the mean free path l. If a magnetic field is switched on the path of a particle in an infinite sample will be a section of a cycloid l long, and *during the mean free time the particle will travel a distance less than l in the direction of the field* **E**, i.e.

$$l_x \cong l\cos\varphi \cong l\left(1-\frac{\varphi^2}{2}\right)\cong l\left(1-\frac{\mu^2B^2}{2}\right). \tag{41.22}$$

Since the distance covered by the particle during the time τ in the direction of the field **E** decreases, this is tantamount to a decrease of drift velocity, or mobility and, for this reason, of conductivity, too, i.e. *there should be an increase in the resistance.* Obviously,

$$\frac{\sigma_0-\sigma}{\sigma_0}=\frac{l_0-l}{l_0}=\frac{\mu^2B^2}{2}, \tag{41.23}$$

or

$$\frac{\rho-\rho_0}{\rho_0}=\frac{\mu^2B^2}{2}. \tag{41.24}$$

If the statistical spread in the free times (and paths) is taken into account, it may be obtained

$$\frac{\Delta\rho}{\rho_0}=\mu^2B^2. \tag{41.25}$$

Thus, *in a magnetic field the resistance increases.*

In a finite semiconductor sample the Hall field compensates the action of the magnetic field with the result that the charge carriers move in straight lines; therefore, from this point of view there should be no magnetoresistance. Actually, however, it remains in this case as well, because *the Hall field compensates the action of the magnetic field only on the average, as if all the charge carriers were travelling with the same (drift) velocity.* But the velocities of the electrons (and holes) are not the same, and therefore *the particles with greater velocities are more affected by the magnetic field than by the Hall field*, and *the slower particles, on the contrary, are deflected by the prevailing Hall field.* The spread in particle velocities leads to a *decrease in the contribution to conductivity of charge carriers with high and low velocities*, and thus to an increase in resistance, although to a far lesser degree than in an infinite semiconductor. The magnetoresistive effect is sensitive to the shape of the sample. An infinite sample may be simulated by a disc (the Corbino disc). Since in this disc the current flows radially, the charge carriers are deflected by the magnetic field in a direction perpendicular to the radius with the result that their separation and accumulation do not take place, and the Hall field does not appear.

If the magnetic field is directed along **j** there should be no change in resistance. However, some substances exhibit magnetoresistance, the explanation for this being the intricate shape of their constant-energy surfaces. In some substances the application of a magnetic field results in an increase in conductivity, the effect being termed *negative magnetoresistance.*

3. Ettingshausen effect. Although the Lorentz force and the Hall field on the average compensate each other, *the spread in charge carriers velocities results in a situation when some of them (the "hot" and the "cold") are deflected to the opposite faces of the sample.* The electrons attain thermodynamic equilibrium with the lattice as a result of collisions with its atoms. If they give up energy, the semiconductor will be heated; if they receive energy, it will be cooled. This leads to a *temperature gradient being established in the direction perpendicular both to* **B** *and the current density* **j** *(the Ettingshausen effect).* The effect is described by the Ettingshausen coefficient A^E:

$$\nabla T^E = -A^E[\mathbf{Bj}] = A^E[\mathbf{jB}], \qquad (41.26)$$

or.

$$\frac{\partial T^E}{\partial z} = A^E B_y j_x; \quad A^E = \frac{\nabla_z T^E}{B_y j_x}. \qquad (41.27)$$

As the magnetic field (or the current) is reversed the sign of $\nabla_z T^E$ changes in the same way as that of the Hall field. *The Hall and the Ettingshausen effects which depend on the direction of* **B** *are termed odd effects.*

4. Nernst, or longitudinal galvanothermomagnetic effect. The essence of the effect is that *a temperature gradient, which is independent of the direction of the magnetic field but reverses its* *sign when the current is reversed, is established along the current* **j**. The cause of the Nernst effect is the decrease of the current of "hot" and "cold" electrons in the direction of their flow. *The Gauss and Nernst effects are even.*

In conclusion of the section it should be pointed out that *the galvanomagnetic effects are subdivided into the adiabatic and the isothermal. A phenomenon is termed adiabatic when there is no exchange of heat between the sample and the ambient. The term isothermal applies when energy exchange between the sample and the ambient prevents the establishment of a temperature gradient in the direction perpendicular to the magnetic field* **B** *and the current* **j**.

Summary of Sec. 41

1. Phenomena observed when a current carrying semiconductor is placed in a magnetic field are termed galvanomagnetic. They are termed adiabatic when there is no heat exchange between the ambient and the sample through its side surface. If as a result of energy exchange with the ambient the temperature of the semiconductor in a plane perpendicular to the current is everywhere the same, the effect should be termed isothermal.

2. The essence of the Hall effect is the appearance of a transverse electric field $\mathbf{E}^H$:

$$\mathbf{E}^H = R[\mathbf{Bj}] = -R[\mathbf{jB}]. \tag{41.1s}$$

The quantity

$$R = -\frac{\mathbf{E}^H}{|\mathbf{jB}|} \tag{41.2s}$$

is termed the Hall coefficient.

3. The term magnetoresistance or magnetoresistive effect applies to the change in the resistance of a semiconductor due to the magnetic field. The magnetoresistance coefficient H defined by the relation

$$H = \frac{1}{B^2}\frac{\rho-\rho_0}{\rho_0} \tag{41.3s}$$

serves to describe the effect.

4. The essence of the Ettingshausen, or the transverse galvanothermomagnetic effect is that a transverse temperature gradient is established:

$$\nabla T^E = A^E[\mathbf{jB}]. \tag{41.4s}$$

A^E is termed the Ettingshausen coefficient. It may be defined by the relation

$$A^E = \frac{\nabla_z T^E}{B_y j_x}. \tag{41.5s}$$

5. The Nernst, or the longitudinal galvanothermomagnetic effect consists in the appearance of a longitudinal (relative to the current) temperature gradient:

$$\frac{\partial T^N}{\partial x} = A^N B_y^2 j_x. \qquad (41.6s)$$

6. When the direction of the current is reversed all the galvanomagnetic effects change sign, i.e. the direction of the fields and of the temperature gradients is reversed.

7. The signs of the Hall and Ettingshausen effects change with the direction of the magnetic field (odd, or transverse effects), while the signs of the magnetoresistive and the Nernst effects remain unaltered (even, or longitudinal effects).

8. For given direction of electric current the electrons and holes move (drift) in opposite directions, their deflection in a magnetic field caused by the drift velocity, is, however, the same. Because of this the signs of the Hall fields are opposite in n- and p-type semiconductors. The Ettingshausen, the Nernst and the magnetoresistive effects are indifferent to the sign of the charge carriers.

9. The Ettingshausen effect is necessarily adiabatic. The other effects may be both adiabatic and isothermal.

10. The sign of the Ettingshausen and the Nernst effects depends on the scattering mechanism, since

$$\varphi = \omega_c \tau = \omega_c \tau_0 E^p, \qquad (41.7s)$$

and (for small φ)

$$\frac{\Delta l}{l} \cong \frac{\varphi^2}{2} = \frac{\omega_c^2 \tau_0^2}{2} E^{2p}. \qquad (41.8s)$$

Therefore,

$$\frac{d}{dE}\left(\frac{\Delta l}{l}\right) = 2p\omega_c^2 \tau_0^2 E^{2p-1}. \qquad (41.9s)$$

For $p=0$ the relative change in the mean free path of the charge carriers is independent of energy; therefore, the velocity spread should not affect the composition of the current in the presence of a magnetic field, and this should result in the absence of the Ettingshausen and the Nernst effects. For $p > \frac{1}{2}$ the relative part played by "hot" electrons is diminished, and the temperature of the semiconductor should drop in the direction of the charge carrier motion; $\nabla_x T$ should coincide in direction with the current j_x in an electron-type semiconductor and be opposite to it in the case of a hole-type semiconductor. For $p < 0$ the contribution of "hot" electrons to the current is increased with the resultant change in the sign of the temperature gradient, as compared to the former case.

42. HALL EFFECT IN EXTRINSIC CONDUCTIVITY RANGE

Now we will use the kinetic equation to describe the Hall effect. Suppose an electric **E** and a magnetic **B** fields are applied to a substance. We will assume the semiconductor to be homogeneous and isothermal: $\nabla F = \nabla T = 0$.

In compliance with (38.21) the expression for current density may be written in the form

$$\mathbf{j} = e^2 K'_{11}\mathbf{E} + \frac{e^3}{m^*} K'_{12}[\mathbf{EB}] + \frac{e^4}{m^{*2}} K'_{13}\mathbf{B}\,(\mathbf{EB}). \tag{42.1}$$

We will confine ourselves to the case of transverse fields: $(\mathbf{EB}) = 0$. In this case (42.1) will be simplified:

$$\mathbf{j} = e^2 K'_{11}\mathbf{E} + \frac{e^3}{m^*} K'_{12}[\mathbf{EB}]. \tag{42.2}$$

To solve the equation (42.2) additional conditions, analogous to boundary conditions of differential equations, should be available.

To begin with consider an infinite semiconductor. Let the fields $\mathbf{E} = (E_x, 0, 0)$ and $\mathbf{B} = (0, B_y, 0) = (0, B, 0)$ be applied to it. A current density $\mathbf{j}$ is established in this case with non-zero components j_x and j_z:

$$j_x = e^2 K'_{11} E_x; \quad j_y = 0; \quad j_z = \frac{e^3}{m^*} K'_{12} B E_x. \tag{42.3}$$

The j_z component is due to the Lorentz force which acts on the charge carriers moving with a non-zero drift velocity. The scattering mechanism makes itself manifest through the action of the kinetic coefficient K'_{12}, and the drift velocity—through the dependence of j_z on E_x. The angle φ between the direction of the full current $\mathbf{j}$ and the electric field $\mathbf{E}$ will be termed the Hall angle. It may be determined from the condition

$$\tan\varphi = \frac{j_z}{j_x} = \frac{e}{m^*}\frac{K'_{12}}{K'_{11}} B = \mu^H B. \tag{42.4}$$

The quantity μ^H with the dimensionality of mobility determines the Hall angle; it is termed Hall's mobility

$$\mu^H = \frac{e}{m^*}\frac{K'_{12}}{K'_{11}}. \tag{42.5}$$

Its sign depends on the sign of the charge carriers. For $E_x > 0$ and $B_y > 0$ the component $j_z > 0$ for holes and $j_z < 0$ for electrons. The condition $j_z \gtrless 0$ means that $\mathbf{j}$ rotates *counterclockwise or clockwise*, respectively.

In a finite semiconductor the phenomenon takes a different course. Suppose the sample is of the form of a parallelepiped with its edges directed along the co-ordinate axis. Evidently, the current will flow only in the x-direction because such is the sample electric circuit. The j_y and j_z components of the current density are zero. But this is possible only if, besides E_x, there will be other non-zero components of the electric field $\mathbf{E}$. Indeed, should we set $\mathbf{E} = (E_x, 0, 0)$, and simultaneously $\mathbf{B} = (0, B, 0)$ and $\mathbf{j} = (j_x, 0, 0)$, it would follow from (42.3) that $j_z = 0$ requires $E_x = 0$ and $j_x = 0$. This means that the three above conditions as applied to a finite semiconductor sample are *incompatible*. Therefore we will presume that

$$\mathbf{E} = (E_x, E_y, E_z); \quad \mathbf{B} = (0, B, 0); \quad \mathbf{j} = (j_x, 0, 0);$$
$$[\mathbf{EB}] = (-E_zB, 0, E_xB).$$

Write the expression (42.2) for the current:

$$j_x = e^2K'_{11}E_x - \frac{e^3}{m^*}K'_{12}BE_z \neq 0; \tag{42.6}$$

$$j_y = e^2K'_{11}E_y = 0; \quad E_y = 0; \tag{42.7}$$

$$j_z = e^2K'_{11}E_z + \frac{e^3}{m^*}K'_{12}BE_x = 0. \tag{42.8}$$

Find E_z from (42.8):

$$E_z = -\frac{e}{m^*}\frac{K'_{12}}{K'_{11}}BE_x = -\mu^H BE_x. \tag{42.9}$$

The appearance of the field E_z is a manifestation of the Hall effect, i.e. E_z *is the Hall electric field*. The total field $\mathbf{E}$ rotates through the angle φ determined from the expression

$$\tan\varphi = \frac{E_z}{E_x} = -\mu^H B. \tag{42.10}$$

The condition (42.10) is the same as (42.4) with the exception of the sign. This is quite natural, since when $\mathbf{j}$ rotates clockwise relative to $\mathbf{E}$, the vector $\mathbf{E}$ should rotate counterclockwise relative to $\mathbf{j}$. *The angle φ as determined from* (42.10) *will be termed Hall's angle, too*. Express now j_x in terms of E_x, substituting into (42.6) the expression (42.9) for E_z:

$$j_x = e^2K'_{11}E_x + \frac{e^4}{m^{*2}}\frac{K'^2_{12}}{K'_{11}}B^2E_x = \sigma_B E_x, \tag{42.11}$$

where

$$\sigma_B = e^2K'_{11} + \frac{e^4}{m^{*2}}\frac{K'^2_{12}B^2}{K'_{11}} = e^2K'_{11}\left[1 + (\mu^H)^2 B^2\right] \tag{42.12}$$

is the specific electrical conductance in the direction of the field E_x in the presence of the magnetic field.

Introduce now the Hall coefficient.' R, *according to* (42.12), *is introduced by the condition*

$$R = -\frac{\mathbf{E}^H}{|\mathbf{jB}|}. \tag{42.13}$$

But

$$\mathbf{E}^H = (0, 0, E_z); \quad [\mathbf{jB}] = (0, 0, j_x B), \tag{42.14}$$

therefore, it follows from (42.13) and (42.14) that

$$E_z = -Rj_x B; \quad R = -\frac{E_z}{j_x B_y}. \tag{42.15}$$

On the other hand, from (42.9) and (42.11) it follows

$$E_z = -\mu^H B E_x = -\mu^H B \frac{j_x}{\sigma_B}. \tag{42.16}$$

Comparing (42.15) and (42.16) we may write

$$R = \frac{\mu^H}{\sigma_B}, \tag{42.17}$$

or

$$\mu^H = R\sigma_B. \tag{42.18}$$

Substituting μ^H from (42.5) and σ_B from (42.12) we obtain

$$R = \frac{\mu^H}{\sigma_B} = \frac{\mu^H}{e^2 K'_{11} [1 + (\mu^H)^2 B^2]} = \frac{\frac{e}{m^*} \frac{K'_{12}}{K'_{11}}}{e^2 K'_{11} \left[1 + \frac{e^2}{m^{*2}} \left(\frac{K'_{12}}{K'_{11}} \right)^2 B^2 \right]}. \tag{42.19}$$

The expression (42.19) *for R is valid for all magnetic fields (if the changes in the energy spectrum are ignored) provided the conductivity is due to charge carriers of one type only, i.e. for extrinsic conductiv ty range.* Since

$$K'_{11} = \frac{n}{m^*} \left\langle \frac{\tau}{1 + \mu^2 B^2} \right\rangle \tag{42.20}$$

and

$$K'_{12} = \frac{n}{m^*} \left\langle \frac{\tau^2}{1 + \mu^2 B^2} \right\rangle, \tag{42.21}$$

it follows that

$$R=\frac{\dfrac{e}{m^*}\dfrac{\left\langle\dfrac{\tau^2}{1+\mu^2B^2}\right\rangle}{\left\langle\dfrac{\tau}{1+\mu^2B^2}\right\rangle}}{\dfrac{e^2n}{m^*}\left\langle\dfrac{\tau}{1+\mu^2B^2}\right\rangle\left[1+\dfrac{e^2}{m^{*2}}\dfrac{\left\langle\dfrac{\tau^2}{1+\mu^2B^2}\right\rangle^2}{\left\langle\dfrac{\tau}{1+\mu^2B^2}\right\rangle^2}B^2\right]}. \qquad (42.22)$$

This expression shows that *the Hall coefficient R depends in a complex way on magnetic field; it is inversely proportional to charge carrier concentration, and its sign coincides with the sign of the charge carriers.*

Consider two extreme cases—one of weak, and the other of strong fields.

Taking into account that $\mu=\mu(E)=\frac{e\tau(E)}{m^*}$ we re-write (42.22) in the following form:

$$R=\frac{1}{en}\,\frac{\left\langle\dfrac{\mu^2}{1+\mu^2B^2}\right\rangle}{\left\langle\dfrac{\mu}{1+\mu^2B^2}\right\rangle^2+\left\langle\dfrac{\mu^2}{1+\mu^2B^2}\right\rangle^2B^2}. \qquad (42.23)$$

The condition for *a weak magnetic field* is

$$\mu^2B^2\ll 1. \qquad (42.24)$$

For *a strong magnetic field* the inequality is reversed:

$$\mu^2B^2\gg 1. \qquad (42.25)$$

The conditions (42.24) and (42.25) are less exacting than (41.5) and (41.6) and, therefore, more easily fulfilled.

1. **Weak magnetic field.** In a weak magnetic field

$$K'_{11}=\frac{n}{m^*}\left\langle\frac{\tau}{1+\mu^2B^2}\right\rangle\cong\frac{n}{m^*}\langle\tau\rangle=K_{11};$$
$$K'_{12}\cong\frac{n}{m^*}\langle\tau^2\rangle=K_{12}. \qquad (42.26)$$

Besides,

$$\frac{e^2}{m^{*2}}\left(\frac{K'_{12}}{K'_{11}}\right)^2B^2\cong\frac{e^2}{m^{*2}}\left(\frac{K_{12}}{K_{11}}\right)^2B^2=\frac{e^2}{m^{*2}}\left(\frac{\langle\tau^2\rangle}{\langle\tau\rangle}\right)^2B^2=$$
$$=\frac{\langle\tau^2\rangle^2}{\langle\tau\rangle^4}\langle\mu^2B^2\rangle\ll 1, \qquad (42.27)$$

since $\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}$ is of the order of unity. Taking into account (42.27) we may write R in the form

$$R=\frac{\frac{e}{m^*}\frac{K_{12}}{K_{11}}}{e^2K_{11}}=\frac{1}{em^*}\frac{K_{12}}{K_{11}^2}=\frac{1}{en}\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}=\frac{A}{en}. \qquad (42.28)$$

The expression (42.28) for R in a weak magnetic field is independent of B and coincides, to a constant A (the Hall factor)

$$A=\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}, \qquad (42.29)$$

with the expression (41.17) which was obtained on the basis of elementary notions concerning the Hall effect. *The quantity* A *is dependent on the scattering mechanism. For non-degenerate semiconductors and for* $\tau=\tau_0E^p=\tau_0'x^p$

$$A=\frac{\dfrac{\Gamma\left(\frac{5}{2}+2p\right)}{\Gamma\left(\frac{5}{2}\right)}}{\left[\dfrac{\Gamma\left(\frac{5}{2}+p\right)}{\Gamma\left(\frac{5}{2}\right)}\right]^2}=\frac{\Gamma\left(\frac{5}{2}+2p\right)\Gamma\left(\frac{5}{2}\right)}{\left[\Gamma\left(\frac{5}{2}+p\right)\right]^2}. \qquad (42.30)$$

Table 11 shows the values of the Hall factor A for various p's.

Table 11

p	0	1/2	1	3/2	$-1/2$	-1	2
A	1	$\frac{45\pi}{128}=1.13$	$\frac{7}{5}=1.40$	$\frac{315\pi}{512}=1.93$	$\frac{3\pi}{8}=1.18$	$\frac{27\pi}{16}=5.30$	$\frac{99}{35}=2.83$

It will be demonstrated in the following chapter that *for scattering by impurity ions* $p=3/2$, *and* $A=\frac{315\pi}{512}=1.93$; *for scattering by lattice vibrations* $p=-1/2$, *and* $A=\frac{3\pi}{8}=1.18$. Since the choice of the scattering mechanism is determined by the temperature range, it may be said that *in the low-temperature range when the scattering by impurity atoms prevails* A *should be put equal to* 1.93. *In the temperature range where the scattering by lattice vibrations is the dominant factor*, A = 1.18. The Hall mobility in a weak magnetic field is

$$\mu^H = \frac{e}{m^*}\frac{K_{12}}{K_{11}} = \frac{e}{m^*}\frac{\langle\tau^2\rangle}{\langle\tau\rangle}. \qquad (42.31)$$

Since the expression for drift mobility in a weak magnetic field is

$$\mu_d = \frac{e}{m^*}\langle\tau\rangle, \qquad (42.32)$$

it follows

$$\mu^H = \frac{e\langle\tau\rangle}{m^*}\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2} = A\mu_d, \qquad (42.33)$$

i.e. *the Hall mobility which determines the Hall angle is proportional to drift mobility.*

Consider the quantity σ_B. According to (42.27) and (42.12) it may be written in the form (for weak field!):

$$\sigma_B = e^2K'_{11}[1 + (\mu^H)^2B^2] \cong e^2K_{11} = \sigma_0. \qquad (42.34)$$

It follows from (42.18) that

$$\mu^H = R\sigma_0. \qquad (42.35)$$

Taking the product of R and σ_0 both of which have been measured experimentally we obtain μ^H. If the scattering mechanism is known $\mu_d = \mu^H/A$ may be determined from μ^H, and carrier concentration and their sign—from R. *This makes the Hall effect one of the most important effects for the study of semiconductors.*

For degenerate semiconductors and metals:

$$A = \frac{\langle\tau^2\rangle}{\langle\tau\rangle^2} = \frac{\tau^2(F)}{[\tau(F)]^2} = 1, \qquad (42.36)$$

therefore

$$\mu^H = \mu_d; \quad R = \frac{1}{en}. \qquad (42.37)$$

2. Strong magnetic fields. In strong fields:

$$\frac{K'_{12}}{K'_{11}} = \frac{\left\langle\frac{\tau^2}{1+\mu^2B^2}\right\rangle}{\left\langle\frac{\tau}{1+\mu^2B^2}\right\rangle} \cong \frac{\left\langle\tau^2\left(\frac{e^2\tau^2}{m^{*2}}\right)^{-1}\cdot\right\rangle\frac{1}{B^2}}{\left\langle\tau\left(\frac{e^2\tau^2}{m^{*2}}\right)^{-1}\right\rangle\frac{1}{B^2}} = \frac{1}{\left\langle\frac{1}{\tau}\right\rangle} = \langle\tau^{-1}\rangle^{-1} \qquad (42.38)$$

and

$$\left(\frac{K'_{12}}{K'_{11}}\right)^2 \cong \langle\tau^{-1}\rangle^{-2}. \qquad (42.39)$$

Taking into account (42.39) re-write the expression (42.19) for the case of a strong field:

$$R \cong \frac{\langle \tau^{-1} \rangle^{-1}}{en \cdot \langle \tau^{-1} \rangle \left(\frac{m^{*2}}{e^2} \cdot \frac{1}{B^2} + \langle \tau^{-1} \rangle^{-2} \right)} . \qquad (42.40)$$

For fields of unlimited intensity the first addend of the denominator in (42.40) may be smaller than the second, therefore

$$R = \frac{1}{en}, \qquad (42.41)$$

i.e. *in a strong magnetic field, as in the case of a degenerate semiconductor, the Hall coefficient is independent of the scattering mechanism* (A = 1). Comparing the expressions (42.41) and (42.28) for R in strong and weak fields we see that R depends on B: as the field increases A diminishes from $\frac{A}{en}$ to $\frac{1}{en}$. The transition takes place in the range of field intensities for which

$$\mu^2 B^2 \cong 1. \qquad (42.42)$$

The quantity A may be easily obtained as the ratio of R's in weak—R_0—and strong—R_∞—magnetic fields:

$$A = \frac{R_0}{R_\infty}. \qquad (42.43)$$

Note in conclusion that the concept of strong and weak fields depends on the mobility of carriers of specific energy. But for $\tau = \tau_0 E^p$ and $p > 0$ the condition $\mu^2 B^2 \ll 1$ cannot formally be satisfied for any of the possible energy values. Actually, however, this is not the case. Both the number of the high-energy charge carriers and their contribution to the averaged relaxation times are negligible. *Therefore, the conditions* (42.24) *and* (42.25) *should be understood in the sense that they will be satisfied if the drift (or the Hall) mobility is taken for* μ, *i.e. the fields should be assessed from the conditions:*

$$\mu_d^2 B^2 \ll 1; \quad \mu_d^2 B^2 \gg 1. \qquad (42.44)$$

Consider an example. In InSb $\mu_d \cong 8$ m^2 (V·s)$^{-1}$, therefore, a weak field should satisfy the condition $B^2 \ll 1/64 = 0.015$, or $B < 0.1$ T = 1000 Gs. In p-Si $\mu_d \cong 0.05$ m^2/(V.s), and all practically attainable fields are weak.

Summary of Sec. 42

1. The equation for current density

$$\mathbf{j}=e^2K'_{11}\mathbf{E}+\frac{e^3}{m^*}K'_{12}[\mathbf{EB}] \tag{42.1s}$$

with "boundary" conditions

$$\mathbf{j}=(j_x,\ 0,\ 0) \tag{42.2s}$$

leads to an expression for a transverse field E_z, the Hall field:

$$E_z=-\frac{\mu^H}{\sigma_B}B_y j_x=-RB_y j_x, \tag{42.3s}$$

whence follows the expression for the Hall coefficient R

$$R=\frac{\mu^H}{\sigma_B}. \tag{42.4s}$$

2. The Hall mobility μ^H is determined by the Hall field intensity, by the field intensity E_x responsible for the current, and by the magnetic induction $B=B_y$:

$$\mu^H=\frac{E_z}{B_y E_x}; \tag{42.5s}$$

the Hall angle is determined by the Hall mobility μ^H:

$$\tan\varphi=\frac{E_z}{E_x}=\mu^H B. \tag{42.6s}$$

The Hall mobility is expressed in the kinetic coefficients:

$$\mu^H=\frac{e}{m^*}\frac{K'_{12}}{K'_{11}}=\frac{e}{m^*}\frac{\left\langle\frac{\tau^2}{1+\mu^2B^2}\right\rangle}{\left\langle\frac{\tau}{1+\mu^2B^2}\right\rangle}. \tag{42.7s}$$

3. In a weak magnetic field ($\mu_d^2B^2\ll 1$)

$$\sigma_B\cong\sigma_0=en\mu_d;\quad \mu^H=\frac{e}{m^*}\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}=A\mu_d;$$
$$R=\frac{A}{en}. \tag{42.8s}$$

4. In a strong magnetic field

$$R=\frac{1}{en};\quad \mu^H=\frac{e}{m^*}\langle\tau^{-1}\rangle^{-1}, \tag{42.9s}$$

i.e. the Hall coefficient is independent of the scattering mechanism.

5. The scattering mechanism affects the value of R through the factor A; in a weak field the Hall factor

$$\mathrm{A}=\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}. \tag{42.10s}$$

This is equal to $\frac{3\pi}{8}$ for scattering by lattice vibrations, to $\frac{315\pi}{512}$ for scattering by impurity ions, and to unity in degenerate semiconductors and metals. In a strong field $\mathrm{A}=1$.

6. Charge carrier concentration is

$$n=\frac{\mathrm{A}}{eR}. \tag{42.11s}$$

43. HALL EFFECT IN A SUBSTANCE WITH SEVERAL TYPES OF CHARGE CARRIERS

In this section we will discuss *the Hall effect in a uniform semiconductor under isothermal conditions presuming that it contains several types of charge carriers.* This may be an intrinsic semiconductor, a semiconductor in the mixed conductivity range, or a semiconductor in the extrinsic range provided it contains charge carriers of different effective masses, for instance p-Si or p-Ge which contain light and heavy holes.

For the charge carriers of the type α the usual expression for the current may be written as

$$\mathbf{j}_\alpha=e_\alpha^2K'_{11(\alpha)}\mathbf{E}+\frac{e_\alpha^3}{m_\alpha^*}K'_{12(\alpha)}[\mathbf{EB}]. \tag{43.1}$$

The total current

$$\mathbf{j}=\sum_\alpha \mathbf{j}_\alpha=\Big(\sum_\alpha e_\alpha^2K'_{11(\alpha)}\Big)\mathbf{E}+\Big(\sum_\alpha\frac{e_\alpha^3}{m_\alpha^*}K'_{12(\alpha)}\Big)[\mathbf{EB}]. \tag{43.2}$$

All the expressions of the preceding section should be valid if corresponding sums are substituted for K'_{11} and K'_{12}. Consider the solution of the equation (43.2) for a finite semiconductor sample:

$$\mathbf{B}=(0,\ B,\ 0);\quad \mathbf{E}=(E_x,\ E_y,\ E_z);\quad [\mathbf{EB}]=(-E_zB,\ 0,\ E_xB);$$
$$\mathbf{j}=(j_x,\ 0,\ 0). \tag{43.3}$$

The Hall coefficient R is determined by the condition (42.15). Find j_x and j_z:

$$j_x=\Big(\sum_\alpha e_\alpha^2K'_{11(\alpha)}\Big)E_x-\Big(\sum_\alpha\frac{e_\alpha^3}{m_\alpha^*}K'_{12(\alpha)}\Big)BE_z; \tag{43.4}$$

$$j_y=0,\quad E_y=0; \tag{43.5}$$

$$j_z=\Big(\sum_\alpha e_\alpha^2K'_{11(\alpha)}\Big)E_z+\Big(\sum_\alpha\frac{e_\alpha^3}{m_\alpha^*}K'_{12(\alpha)}\Big)BE_x=0. \tag{43.6}$$

Determine E_z from (43.6):

$$E_z = -\frac{\left(\sum_\alpha \frac{e_\alpha^3}{m_\alpha^*} K'_{12(\alpha)}\right) BE_x}{\left(\sum_\alpha e_\alpha^2 K'_{11(\alpha)}\right)}. \tag{43.7}$$

This expression may be represented in the same form as (42.9) if the Hall mobility μ^H is defined likewise:

$$\tan\varphi = \frac{E_z}{E_x} = -\mu^H B. \tag{43.8}$$

Then

$$\mu^H = \frac{\left(\sum_\alpha \frac{e_\alpha^3}{m_\alpha^*} K'_{12(\alpha)}\right)}{\left(\sum_\alpha e_\alpha^2 K'_{11(\alpha)}\right)^2}, \tag{43.9}$$

and

$$E_z = -\mu^H B E_x. \tag{43.10}$$

Express E_z in terms of j_x

$$j_x = \left(\sum_\alpha e_\alpha^2 K'_{11(\alpha)}\right) E_x + \frac{\left(\sum_\alpha \frac{e_\alpha^3}{m_\alpha^*} K'_{12(\alpha)}\right)^2}{\left(\sum_\alpha e_\alpha^2 K'_{11(\alpha)}\right)} B^2 E_x = \sigma_B E_x, \tag{43.11}$$

where

$$\sigma_B = \left(\sum_\alpha e_\alpha^2 K'_{11(\alpha)}\right)\{1 + (\mu^H)^2 B^2\} \tag{43.12}$$

is the conductivity in the direction of the current in the presence of the magnetic field. From (42.15), (43.10) and (43.11) we may write

$$R = -\frac{E_z}{Bj_x} = -\frac{E_z}{B\sigma_B E_x} = \frac{\mu^H B E_x}{\sigma_B B E_x} = \frac{\mu^H}{\sigma_B}, \tag{43.13}$$

which is quite analogous to (42.17). However, the expressions for σ_B and μ^H are in this case different. Substituting the expressions for them (43.12) and (43.9) into (43.13) we obtain

$$R = \frac{\mu^H}{\sigma_B} = \frac{\mu^H}{\left(\sum_\alpha e_\alpha^3 K'_{11(\alpha)}\right)\{1 + (\mu^H)^2 B^2\}} =$$

$$= \frac{\left(\sum_\alpha \frac{e_\alpha^3}{m_\alpha^*} K'_{12(\alpha)}\right)}{\left(\sum_\alpha e_\alpha^2 K'_{11(\alpha)}\right)^2 + \left(\sum_\alpha \frac{e_\alpha^3}{m_\alpha^*} K'_{12(\alpha)}\right)^2 B^2}. \tag{43.14}$$

Substituting the expressions (29.27) for the kinetic coefficients into (43.14) we write:

$$R=\frac{\sum\limits_{\alpha} e_\alpha n_\alpha \left\langle \frac{\mu_\alpha^2}{1+\mu_\alpha^2 B^2}\right\rangle}{\left(\sum\limits_{\alpha} e_\alpha n_\alpha \left\langle \frac{\mu_\alpha}{1+\mu_\alpha^2 B^2}\right\rangle\right)^2+\left(\sum\limits_{\alpha} e_\alpha n_\alpha \left\langle \frac{\mu_\alpha^2}{1+\mu_\alpha^2 B^2}\right\rangle\right)^2 B^2}. \quad (43.15)$$

If there is only one type of charge carriers in the substance, (43.15) reduces to (42.23).

1. Weak magnetic fields. Consider the case of weak fields: $\mu_\alpha^2 B^2 \ll 1$. *The inequality should hold for all types of charge carriers* and primarily for those with the greatest mobility. Neglecting in all the terms $\mu_\alpha^2 B^2$ as compared to unity, and the second addend in the denominator as compared to the first addend, we obtain

$$R=\frac{\sum\limits_{\alpha} e_\alpha n_\alpha (\mu_\alpha^2)}{\left[\sum\limits_{\alpha} e_\alpha n_\alpha \langle \mu_\alpha \rangle\right]^2}; \quad (43.16)$$

but

$$\langle \mu_\alpha \rangle = \frac{e_\alpha}{m_\alpha^*}\langle \tau_\alpha \rangle = \mu_{d\alpha} \quad (43.17)$$

and

$$\langle \mu_\alpha^2 \rangle = \frac{e_\alpha^2}{m_\alpha^{*2}}\langle \tau_\alpha^2 \rangle = \left(\frac{e_\alpha \langle \tau_\alpha \rangle}{m_\alpha^*}\right)^2 \frac{\langle \tau_\alpha^2 \rangle}{\langle \tau_\alpha \rangle^2} = \mu_{d\alpha}^2 A_\alpha, \quad (43.18)$$

therefore,

$$R=\frac{\sum\limits_{\alpha} e_\alpha n_\alpha A_\alpha \mu_{d\alpha}^2}{\left(\sum\limits_{\alpha} e_\alpha n_\alpha \mu_{d\alpha}\right)^2}=\frac{\sum\limits_{\alpha} \sigma_\alpha \mu_\alpha^H}{\left(\sum\limits_{\alpha} \sigma_\alpha\right)^2}=\frac{\mu^H}{\sigma}, \quad (43.19)$$

where

$$\sigma=\sum_{\alpha} \sigma_\alpha; \quad \mu^H=\frac{\sum\limits_{\alpha} \mu_\alpha^H \sigma_\alpha}{\sum\limits_{\alpha} \sigma_\alpha}. \quad (43.20)$$

Consider the expression (43.19) for two types of charge carriers:

$$R=\frac{e_1 n_1 A_1 \mu_{d1}^2 + e_2 n_2 A_2 \mu_{d2}^2}{(e_1 n_1 \mu_{d1} + e_2 n_2 \mu_{d2})^2}. \quad (43.21)$$

If the conductivity of the semiconductor is of the mixed type, then, assuming

$$n_1 = n; \quad n_2 = p; \quad e_1 = e_n; \quad e_2 = e_p; \quad \mu_{d1} = \mu_n; \quad \mu_{d2} = \mu_p;$$
$$\frac{\mu_n}{\mu_p} = -b; \quad (b > 0),$$

we may write

$$R = \frac{e_n n \mathrm{A}_n \mu_n^2 + e_p p \mathrm{A}_p \mu_p^2}{(e_n n \mu_n + e_p p \mu_p)^2} = \frac{1}{e_p} \frac{\mathrm{A}_p p - \mathrm{A}_n n b^2}{(p + nb)^2}. \tag{43.22}$$

Supposing that $\mathrm{A}_p = \mathrm{A}_n = \mathrm{A}$; $\frac{n}{p} = x$, we obtain

$$R = \frac{\mathrm{A}}{e_p} \frac{p - nb^2}{(p + nb)^2} = \frac{\mathrm{A}}{e_p p} \frac{1 - xb^2}{(1 + xb)^2}. \tag{43.23}$$

Consider three specific cases:
a hole-type semiconductor

$$n = 0; \quad R = \frac{\mathrm{A}}{e_p p}; \quad R > 0; \tag{43.24}$$

an electron-type semiconductor

$$p = 0; \quad R = -\frac{\mathrm{A}}{e_p n} = \frac{\mathrm{A}}{e_n n}; \quad R < 0; \tag{43.25}$$

and an intrinsic semiconductor

$$p = n; \quad x = 1; \quad R = \frac{\mathrm{A}}{e_p p} \frac{1 - b^2}{(1 + b)^2} = \frac{\mathrm{A}}{e_p p} \frac{1 - b}{1 + b}. \tag{43.26}$$

Usually $b > 1$, therefore in an intrinsic semiconductor $R < 0$. For $b = 1$, $R = 0$ and $E_z = 0$; in this case *electrons and holes deflected to the same side do not establish a Hall field since their charges compensate each other*. If, on the other hand, $b \neq 1$, R will not be zero, *its sign being determined by the sign of the carriers with the greater mobility*.

n and p are temperature dependent, and so is R. In the intrinsic temperature range $x = 1$, and R is negative, with its modulus decreasing with the increase in T. If at low temperatures the semiconductor was p-type with $R > 0$, as T increases its Hall coefficient should pass through zero (Fig. 63). This temperature is termed *inversion temperature* T_{inv}. It is determined from the condition

$$R(T_{inv}) = 0 = p - nb^2, \tag{43.27}$$

whence

$$b^2 = \frac{p}{n};\ b = \sqrt{\frac{p}{n}}. \qquad (43.28)$$

If n and p are known, b at T_{inv} may be calculated.

For example, in intermetallic compounds $b \gg 1$, therefore the conductivity may be of hole type ($p \gg bn$), and R may at the same time be negative.

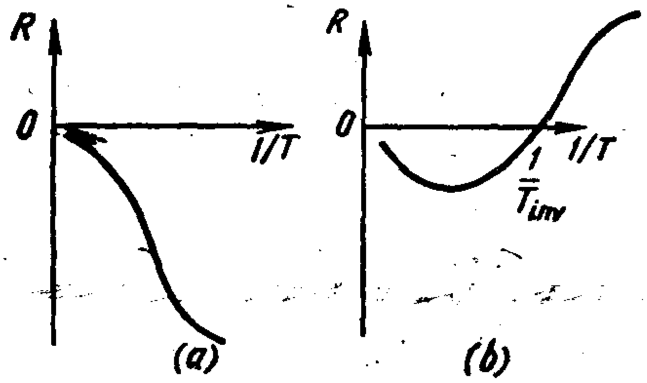


Fig. 63. The dependence of the Hall coefficient on the inverse temperature in an electron-type (a) and a hole-type (b) semiconductors (the conductivity type holds for low temperatures)

Generally, the sign of R coincides with the sign of the quantity

$$1 - xb^2 = 1 - \frac{n}{p}\frac{\mu_n^2}{\mu_p^2}. \qquad (43.29)$$

The Hall coefficient of a semiconductor in the mixed conductivity range is determined by four quantities: n, p, μ_n, μ_p. The same quantities also determine σ.

If the sign of the charge carriers is the same (as in p-Si or p-Ge) R will be of the form

$$R = \frac{1}{e_p}\frac{A_1 p_1 \mu_{p1}^2 + A_2 p_2 \mu_{p2}^2}{(p_1\mu_{p1} + p_2\mu_{p2})^2}. \qquad (43.30)$$

Introducing the notation

$$\frac{\mu_{p2}}{\mu_{p1}} = b;\ \frac{p_2}{p_1} = x;\ A_1 = A_2 = A,$$

we obtain

$$R = \frac{A}{e_p p_1}\frac{1 + xb^2}{(1 + xb)^2}. \qquad (43.31)$$

Both expressions (43.31) and (43.23) show that the *relative contribution of the fast charge carriers to the Hall field is determined by the value of xb^2, while their contribution to conductivity is determined by the value of xb, i.e. charge carriers with greater* *mobility play a greater part in the Hall effect than in conductivity.*

2. Strong magnetic fields. Turn now to the case of strong fields: $\mu_\alpha^2 B^2 \gg 1$. We may write from (43.15)

$$R = \frac{\sum\limits_\alpha e_\alpha n_\alpha}{\left(\sum\limits_\alpha e_\alpha n_\alpha\right)^2 + \left(\sum\limits_\alpha e_\alpha n_\alpha \langle \mu_\alpha^{-1} \rangle\right)^2 \frac{1}{B^2}}. \tag{43.32}$$

For an intrinsic semiconductor the numerator of (43.32) turns zero. The denominator, however, remains non-zero for all finite *B*'s; therefore, for an *intrinsic semiconductor* $R=0$.

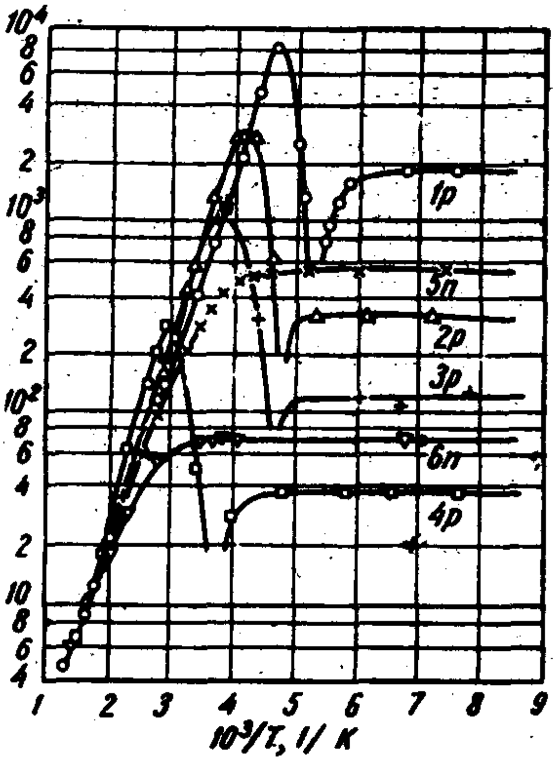


Fig. 64. The dependence of log |*R*| on the inverse temperature in the electron- and the hole-type indium antimonide

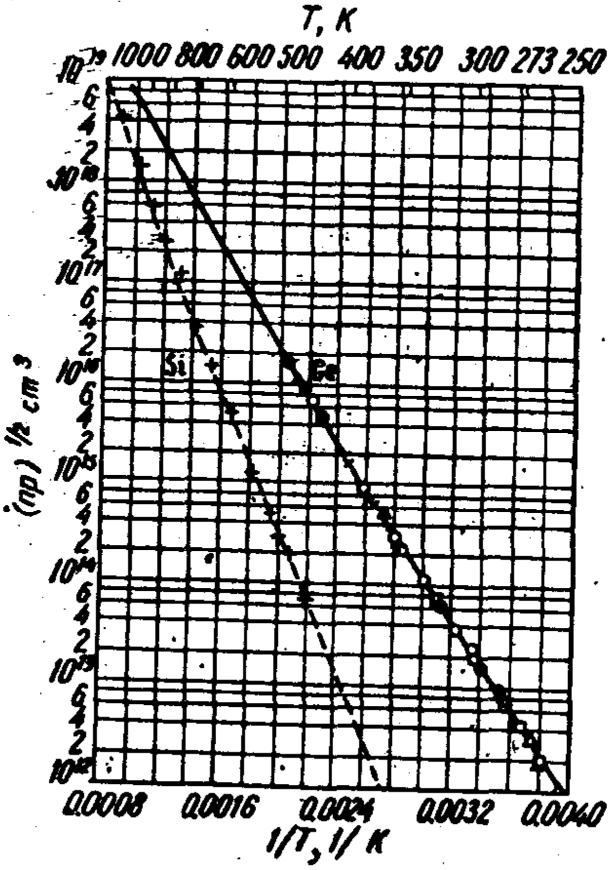


Fig. 65. The dependence of ln n_i on the inverse temperature in germanium and silicon

If the semiconductor is extrinsic,

$$R = \frac{1}{\sum\limits_\alpha e_\alpha n_\alpha} = \frac{1}{\sum\limits_\alpha R_\alpha^{-1}}, \tag{43.33}$$

or

$$R^{-1} = \sum_\alpha R_\alpha^{-1}. \tag{43.34}$$

Figure 64 shows, by way of an example, the dependence of $|R|$ on T^{-1} for several samples of semiconductors with electron (*n*) and hole (p) conductivities. The figure shows that in the low temperature range (from 120° K to 200° K) the charge carrier concentration remains unaltered. This corresponds to the impurity depletion range. The free charge carrier concentration is 3×10^{15} cm^{-3} for $1p$ and $1.5\times$ $\times10^{17}$ cm^{-3} for $4p$. *The inversion temperature is the greater the greater the impurity concentration*, as may be seen from the curves for samples $1p$, $2p$, $3p$ and $4p$. *The temperature of transition to intrinsic conductivity is the lower the lower the impurity concentration.* In the intrinsic conductivity range $\log|R|$ varies linearly with inverse temperature. Figure 65 shows the inverse temperature dependence of n_i for germanium and silicon derived from experimentally determined inverse temperature dependence of the Hall coefficient.

44. MAGNETIC FIELD DEPENDENCE OF HALL COEFFICIENT

The Hall coefficient which enters the expression for $\mathcal{E}^H$, as the calculation carried out on the basis of the kinetic equation shows, depends in a complex way on the magnetic field. The expressions (42.28), (42.41), (43.19) and (43.33) written above are, in effect, valid only in the extreme cases of $B\rightarrow0$ and $B\rightarrow\infty$. Denote these limiting expressions by R_0 and R_∞. Consider now the expressions which should be valid for $B\neq0$ and $B\neq\infty$, respectively. Start with an extrinsic semiconductor with charge carriers of one type, for which, according to (42.23), R is equal to

$$R=\frac{1}{en}\frac{\left\langle\frac{\mu^2}{1+\mu^2B^2}\right\rangle}{\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle^2+\left\langle\frac{\mu^2}{1+\mu^2B^2}\right\rangle^2B^2}. \tag{44.1}$$

Consider first a degenerate semiconductor. Averaging over energy, according to (39.32), is tantamount to assigning values to various quantities which they assume on the Fermi surface. Accordingly, (44.1) takes the form

$$R=\frac{1}{en}\cdot\frac{\frac{\mu^2(F)}{1+\mu^2(F)B^2}}{\frac{\mu^2(F)}{[1+\mu^2(F)B^2]^2}+\frac{\mu^4(F)B^2}{[1+\mu^2(F)B^2]^2}}=\frac{1}{en}, \tag{44.2}$$

i.e. for a degenerate semiconductor the Hall coefficient is independent of magnetic induction over the entire range of its variation.

Consider now a non-degenerate semiconductor. Expand $\frac{1}{1+\mu^2B^2}$ into a series in μ^2B^2:

$$\frac{1}{1+\mu^2B^2}=\sum_m(-1)^m(\mu^2B^2)^m=1-\mu^2B^2+\mu^4B^4-\mu^6B^6+\ldots. \tag{44.3}$$

For a weak field we confine ourselves to the first two terms. The error in this case will not exceed $\mu^4 B^4$. Thus

$$\left\langle \frac{\mu^2}{1+\mu^2B^2}\right\rangle = \langle \mu^2-\mu^4B^2+\ldots\rangle \cong \langle\mu^2\rangle-\langle\mu^4\rangle B^2,$$
$$\left\langle \frac{\mu}{1+\mu^2B^2}\right\rangle = \langle \mu-\mu^3B^2+\ldots\rangle \cong \langle\mu\rangle-\langle\mu^3\rangle B^2. \qquad (44.4)$$

Using (44.4) we derive R:

$$R=\frac{1}{en}\,\frac{\langle\mu^2\rangle-\langle\mu^4\rangle B^2}{[\langle\mu\rangle-\langle\mu^3\rangle B^2]^2+[\langle\mu^2\rangle-\langle\mu^4\rangle B^2]^2B^2}=$$
$$=\frac{1}{en}\,\frac{\langle\mu^2\rangle}{\langle\mu\rangle^2}\left\{1-B^2\left(\frac{\langle\mu^4\rangle}{\langle\mu^2\rangle}-2\frac{\langle\mu^3\rangle}{\langle\mu\rangle}+\frac{\langle\mu^2\rangle^2}{\langle\mu\rangle^2}\right)\right\}. \qquad (44.5)$$

Since $\frac{1}{en}\frac{\langle\mu^2\rangle}{\langle\mu\rangle^2}=R_0$, it follows

$$R=R_B=R_0\left\{1-B^2\left(\frac{\langle\mu^2\rangle^2-2\langle\mu\rangle\langle\mu^3\rangle}{\langle\mu\rangle^2}+\frac{\langle\mu^4\rangle}{\langle\mu^2\rangle}\right)\right\}, \qquad (44.6)$$

i.e. as the field increases, *the Hall coefficient decreases in proportion to* B^2. The coefficient of proportionality depends on the square of mobility and on the scattering mechanism. Indeed, denote

$$\frac{\langle\mu^2\rangle}{\langle\mu\rangle^2}=\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}=\mathrm{A};\quad \frac{\langle\mu^3\rangle}{\langle\mu\rangle^3}=\frac{\langle\tau^3\rangle}{\langle\tau\rangle^3}=\mathrm{C};\quad \frac{\langle\mu^4\rangle}{\langle\mu\rangle^4}=\frac{\langle\tau^4\rangle}{\langle\tau\rangle^4}=\mathrm{D}, \qquad (44.7)$$

then

$$R_B=R_0\left\{1-\mu_d^2B^2\left(\mathrm{A}^2-2\mathrm{C}+\frac{\mathrm{D}}{\mathrm{A}}\right)\right\}. \qquad (44.8)$$

Find the values of the coefficients A, C and D when only one scattering mechanism is active, $\tau=\tau_0E^p$:

$$\left(\mathrm{A}^2-2\mathrm{C}+\frac{\mathrm{D}}{\mathrm{A}}\right)=\left\{\frac{\Gamma\left(\frac{5}{2}+2p\right)\Gamma\left(\frac{5}{2}\right)}{\left[\Gamma\left(\frac{5}{2}+p\right)\right]^2}\right\}^2-$$
$$-2\frac{\Gamma\left(\frac{5}{2}+3p\right)\left[\Gamma\left(\frac{5}{2}\right)\right]^2}{\left[\Gamma\left(\frac{5}{2}+p\right)\right]^3}+\frac{\Gamma\left(\frac{5}{2}+4p\right)\left[\Gamma\left(\frac{5}{2}\right)\right]^2}{\left[\Gamma\left(\frac{5}{2}+p\right)\right]^2\Gamma\left(\frac{5}{2}+2p\right)}. \qquad (44.9)$$

Table 12 shows the values of the coefficient (44.9) for three cases: $p=-1/2$; $p=0$ and $p=3/2$.

Table 12

p	$-1/2$	0	$3/2$
$A^2-2C+\frac{D}{A}$	$\frac{9\pi^2}{64}=1.39=\left(\frac{3\pi}{8}\right)^2$	0	$615=(24.8)^2$

The transition from lattice vibration to impurity ion scattering leads to a sharp change in the value of the coefficient in front of B^2:

$$R_B=R_0(1-1.39\mu_d^2B^2)\quad \text{for } p=-\frac{1}{2},$$
$$R_B=R_0(1-615\mu_d^2B^2)\quad \text{for } p=\frac{3}{2}. \qquad (44.10)$$

Thus, in weak fields the Hall coefficient decreases in proportion to $\mu_d^2B^2$.

Consider now the case of strong fields ($\mu_d^2B^2\gg 1$). The expansion (44.3) is valid in this case, too; however, the series becomes divergent, and one should resort to an expansion in $(\mu_d^2B^2)^{-1}$. Thus, it may be written

$$\frac{1}{1+x}=\frac{1}{x\left(1+\frac{1}{x}\right)}=\frac{1}{x}\sum_{m=0}^{\infty}(-1)^m\left(\frac{1}{x}\right)^m;\quad |x|>1, \qquad (44.11)$$

or

$$\frac{1}{1+\mu^2B^2}=\frac{1}{\mu^2B^2}\sum_m(-1)^m\left(\frac{1}{\mu^2B^2}\right)^m=\frac{1}{\mu^2B^2}\left(1-\frac{1}{\mu^2B^2}+\ldots\right). \qquad (44.12)$$

Taking into account the expansion (44.12) we write

$$\left\langle\frac{\mu^2}{1+\mu^2B^2}\right\rangle=\left\langle\mu^2\left(\frac{1}{\mu^2B^2}-\frac{1}{\mu^4B^4}+\ldots\right)\right\rangle\cong$$
$$\cong\frac{1}{B^4}(1-\langle\mu^{-2}\rangle B^{-2}) \qquad (44.13)$$

and

$$\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle=\left\langle\mu\left(\frac{1}{\mu^2B^2}-\frac{1}{\mu^4B^4}+\ldots\right)\right\rangle=$$
$$=\frac{1}{B^2}(\langle\mu^{-1}\rangle-\langle\mu^{-3}\rangle B^{-2}). \qquad (44.14)$$

Substituting (44.13) and (44.14) into (44.1) and retaining terms of the power of B not below -4, we obtain:

$$R=\frac{1}{en}\,\frac{\frac{1}{B^2}\left(1-\langle\mu^{-2}\rangle B^{-2}\right)}{\langle\mu^{-1}\rangle^2 B^{-4}+\frac{1}{B^4}\left(1-2\langle\mu^{-2}\rangle B^{-2}\right)B^2}=$$

$$=\frac{1}{en}\left\{1+\left(\langle\mu^{-2}\rangle-\langle\mu^{-1}\rangle^2\right)B^{-2}\right\}=R_B. \qquad (44.15)$$

The expression (44.15) may be represented in the form

$$R_B=\left\{1+\frac{\langle\mu^{-2}\rangle-\langle\mu^{-1}\rangle^2}{B^2}\right\}R_\infty. \qquad (44.16)$$

Setting $\tau=\tau_0E^p$ we obtain, for example, for $p=-1/2$

$$R_B=R_\infty\left(1+\frac{1.76}{\mu_0^2B^2}\right), \qquad (44.17)$$

where

$$\mu_0=\frac{e\tau_0'}{m^*},\quad \tau_0'=\tau_0\,(kT)^p.$$

Thus, in strong and in weak fields

$$R_B=\frac{A}{en}\left(1-a\mu_d^2B^2\right)\ \text{and}\ R_B=\frac{1}{en}\left(1+\frac{a'}{\mu_d^2B^2}\right), \qquad (44.18)$$

where a and a' depend on the scattering mechanism. The general trend of $R(B)$ is shown in Fig. 66.

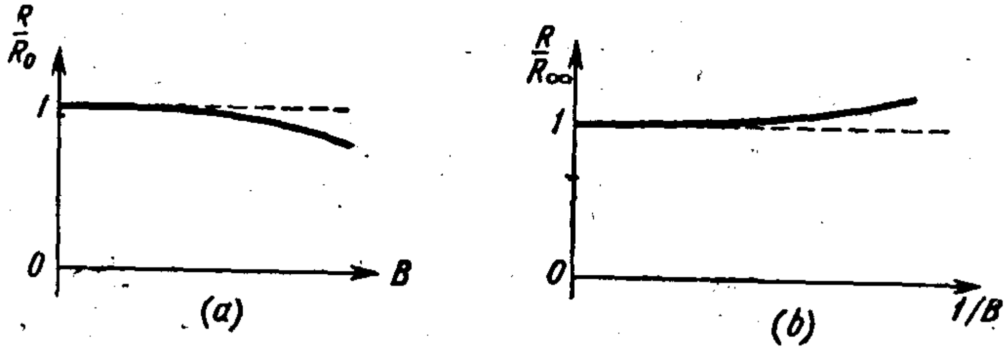


Fig. 66. The dependence of the Hall coefficient on the magnetic field in weak (a) and in strong (b) fields

Investigate now the dependence $R(B)$ for the general case of several types of charge carriers:

$$R=\frac{\sum\limits_\alpha e_\alpha n_\alpha\left\langle\frac{\mu_\alpha^2}{1+\mu_\alpha^2B^2}\right\rangle}{\left[\left(\sum\limits_\alpha e_\alpha n_\alpha\left\langle\frac{\mu_\alpha}{1+\mu_\alpha^2B^2}\right\rangle\right)^2+\left(\sum\limits_\alpha e_\alpha n_\alpha\left\langle\frac{\mu_\alpha^2}{1+\mu_\alpha^2B^2}\right\rangle\right)^2B^2\right]}. \qquad (44.19)$$

For weak fields we obtain, in compliance with (44.4),

$$R=\frac{\sum\limits_\alpha e_\alpha n_\alpha[\langle\mu_\alpha^2\rangle-\langle\mu_\alpha^4\rangle B^2]}{\left\{\left[\sum\limits_\alpha e_\alpha n_\alpha(\langle\mu_\alpha\rangle-\langle\mu_\alpha^3\rangle B^2)\right]^2+\left[\sum\limits_\alpha e_\alpha n_\alpha(\langle\mu_\alpha^2\rangle-\langle\mu_\alpha^4\rangle B^2)\right]^2B^2\right\}}. \tag{44.20}$$

Retain in (44.20) only the terms of the power of B not above 2:

$$R=\frac{\{[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^2\rangle]-[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^4\rangle]B^2\}}{\{[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha\rangle]^2+([\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^2\rangle]^2-2[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha\rangle][\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^3\rangle])B^2\}}=$$

$$=\frac{\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^2\rangle\right]}{\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha\rangle\right]^2}\left\{1-B^2\left[\frac{\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^4\rangle}{\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^2\rangle\right]}+\right.\right.$$

$$\left.\left.+\frac{\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^2\rangle\right]^2-2\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha\rangle\right]\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^3\rangle\right]}{\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha\rangle\right]^2}\right]\right\}, \tag{44.21}$$

where

$$\frac{\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^2\rangle\right]}{\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha\rangle\right]^2}=R_0. \tag{44.22}$$

For $\alpha=1$ the expression (44.21) coincides with (44.6).

For an intrinsic semiconductor $n_1=n=n_2=p$. Let $\langle\mu_1\rangle=$ $=-b\langle\mu_2\rangle$; then

$$R_B=R_0\left\{1-\mu_{pd}^2B^2\left[\frac{D}{A}(1+b^2)+A^2(1-b^2)-2C\frac{1+b^3}{1+b}\right]\right\}. \tag{44.23}$$

For $b\ll 1$

$$R_B=R_0\left\{1-\mu_{pd}^2B^2\left[\frac{D}{A}+A^2-2C\right]\right\}, \tag{44.24}$$

which is in full accord with (44.9) for carriers of one type. For $b\gg 1$

$$R_B=R_0\left\{1-\mu_{nd}^2B^2\left[\frac{D}{A}+A^2-2C\right]\right\}. \tag{44.25}$$

The expressions (44.24) and (44.25) show that *it is the more mobile charge carriers which are mainly responsible for the change in the Hall coefficient in weak magnetic fields.*

The same result was obtained for the case of a semiconductor with charge carriers of one sign but of several types.

Such a situation exists, for instance, in *p*-Ge. As the field increases *R* decreases at the expense of the light holes, since the same field is a strong one for the light holes and a weak one for the heavy holes. In *n*-Ge *R* is, practically, independent of the field in the range from zero to several kilogausses. If the mobilities of the charge carriers of different types differ appreciably, one may visualize the curve *R* (*B*) to have as many inflexions as there are types of charge carriers taking part in conductivity and in the Hall effect. The changes will be the more pronounced the closer are the concentrations of different types of charge carriers to each other.

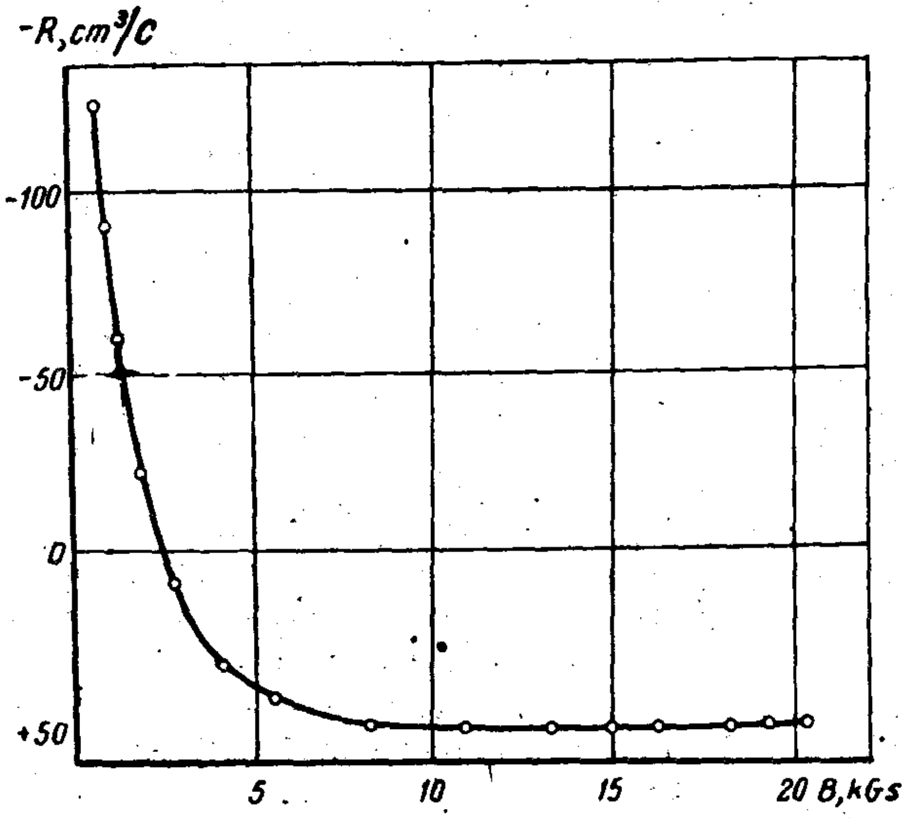


Fig. 67. The field inversion of sign of the Hall coefficient in Cd_xHg_{1-x} Te at 77 K

Let us see how the degeneracy affects the Hall coefficient when different types of charge carriers take part in the conductivity. One may say without analyzing (44.1) that in this case *R* should depend on *B*. To corroborate this statement compare (43.23) and (43.33), whence follows that $R_0 \neq R_\infty$.

In the case of mixed conductivity by degenerate electrons and holes (this is possible in semimetals with overlapping energy bands) the expression (44.1) assumes the form

$$R(B) = \frac{\dfrac{\sigma_1\mu_1}{1+\mu_1^2B^2} + \dfrac{\sigma_2\mu_2}{1+\mu_2^2B^2}}{\left(\dfrac{\sigma_1}{1+\mu_1^2B^2} + \dfrac{\sigma_2}{1+\mu_2^2B^2}\right)^2 + \left(\dfrac{\sigma_1\mu_1B}{1+\mu_1^2B^2} + \dfrac{\sigma_2\mu_2B}{1+\mu_2^2B^2}\right)^2}. \qquad (44.26)$$

Introduce the notation

$$\frac{1+\mu_2^2B^2}{1+\mu_1^2B^2}=f(b,\ B). \tag{44.27}$$

It follows from (44.27) that $f(b,\ 0)=1$ and $f(b,\ \infty)=b^2$. Making use of $f(b,\ B)$ re-write (44.26):

$$R=\frac{1+\mu_1^2B^2}{e_1n_1}\cdot\frac{1+\frac{e_2}{e_1}\,xb^2\cdot\frac{1}{f(b,\ B)}}{\left(1+\frac{xb}{f(b,\ B)}\right)^2+\mu_1^2B^2\left(1+\frac{e_2}{e_1}\,\frac{xb^2}{f(b,\ B)}\right)^2}. \tag{44.28}$$

(44.28) yields (43.23) when $B\longrightarrow 0$ and (43.33) when $B\longrightarrow\infty$. For an intrinsic semimetal we have

$$R=\frac{1+\mu_1^2B^2}{e_pn_i}\cdot\frac{1-\frac{b^2}{f(b,\ B)}}{\left(1+\frac{b}{f(b,\ B)}\right)^2+\mu_p^2B^2\left(1-\frac{b^2}{f(b,\ B)}\right)^2}. \tag{44.29}$$

Although the magnetic field affects the value of the Hall coefficient of degenerate semiconductors as well, the effect, all other conditions being equal, is less pronounced than in non-degenerate semiconductors. This may be seen, for instance, from the relation (44.25): when the Hall field is established predominantly by electrons, electron gas degeneracy results in independence of R from B. From elementary notions of the Hall effect one should expect R to be dependent on B when the magnetic field affects the contribution of charge carriers of different energy to the Hall field in non-degenerate semiconductors with charge carriers of one type, or the contribution of charge carriers of different types in semiconductors with different types of charge carriers independent of degeneracy. In semiconductors and semimetals with high mobility ratios b in weak fields the dominant part may be played by the more mobile charge carriers even if their concentration is small; but as the field increases, their contribution may appreciably diminish, and this makes it possible to observe the change in the sign of the Hall field. Figure 67 shows, by way of an example, the field inversion of the sign of the Hall coefficient of quicksilver telluride in which $b\cong 60$ to 100.

The study of the Hall effect is of the utmost importance for the determination of carrier concentrations and mobilities, and for the investigation of the behaviour of various impurities in semiconductors. The dependence of $\ln|R|$ on T^{-1} enables the ionization energy of an impurity or the width of the forbidden band to be determined; in other words, *the study of the Hall effect yields an appreciable amount of information about the dependence of charge carrier concentration on the ambient or about the composition of the semiconductor.*

Summary of Secs. 43-44

1. The equations for the Hall coefficient R and for the Hall mobility μ^H for semiconductors with charge carriers of several types and for semiconductors with charge carriers of one type are of the same form. The relation between R and μ^H is

$$R\sigma_B=\mu^H. \tag{44.1s}$$

2. The Hall mobility

$$\mu^H=\frac{\sum\limits_\alpha \frac{e_\alpha^3}{m_\alpha^*} K'_{12\,(\alpha)}}{\sum\limits_\alpha e_\alpha^2 K'_{11\,(\alpha)}}=\frac{\sum\limits_\alpha \sigma'_\alpha \mu_\alpha^H}{\sum\limits_\alpha \sigma'_\alpha} \tag{44.2s}$$

is related to the Hall mobility of each type of charge carriers μ_α^H and to the conductivity of charge carriers of each type. The Hall field intensity is determined by the Hall mobility μ^H, the field intensity E_x and magnetic induction B:

$$E_z=-\mu^H BE_x=\tan\varphi E_x. \tag{44.3s}$$

In an intrinsic semiconductor the Hall mobility is

$$\mu^H=\frac{\sigma_n\mu_n^H+\sigma_p\mu_p^H}{\sigma_n+\sigma_p}=\frac{\mu_p^H+b\mu_n^H}{1+b}. \tag{44.4s}$$

If $A_n=A_p$, then

$$\mu^H=\frac{A(1-b^2)\mu_p}{1+b}=(1-b)\mu_p^H. \tag{44.5s}$$

3. The expression for the Hall coefficient is of the form (43.15). In weak fields

$$R=\frac{\mu^H}{\sigma}=\frac{\sum\limits_\alpha \sigma_\alpha \mu_\alpha^H}{\left(\sum\limits_\alpha \sigma_\alpha\right)^2}. \tag{44.6s}$$

For an intrinsic semiconductor

$$R=\frac{A}{e_p p}\,\frac{1-b}{1+b}. \tag{44.7s}$$

In semiconductors with degenerate energy bands the more mobile carriers play a greater part in the Hall field than in conductivity.

4. In strong magnetic fields in an intrinsic semiconductor $R=0$, and in a semiconductor with mixed conductivity

$$R^{-1}=\sum_{\alpha} R_{\alpha}^{-1}=\sum_{\alpha} e_{\alpha} n_{\alpha} . \qquad (44.8s)$$

5. In weak magnetic fields and in strong magnetic fields the Hall coefficient decreases in proportion to $(\mu_d B)^2$ and $(\mu_d B)^{-2}$, respectively.

6. Data obtained in the studies of the Hall effect may be used to determine the temperature dependence of concentration and mobility.

45. MAGNETORESISTIVE EFFECT

Let us apply the kinetic equation to the magnetoresistive effect. Suppose we have a finite parallelepiped, whose edges are directed along the coordinate axes, made of a semiconductor with spherical constant-energy surfaces and with carriers of one type. The current flows along the x-axis. The conductivity in the direction of the field E_x is given by the formula (42.12) repeated here:

$$\sigma_B=e^2K'_{11}+\frac{e^4}{m^{*2}}\frac{(K'_{12})^2}{K'_{11}}B^2, \qquad (45.1)$$

or

$$\sigma_B=en\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle+en\frac{\left\langle\frac{\mu^2}{1+\mu^2B^2}\right\rangle^2 B^2}{\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle}. \qquad (45.2)$$

The magnetoresistance coefficient H is usually introduced by the relation

$$H=\frac{\rho_B-\rho_0}{\rho_0}\cdot\frac{1}{B^2}=\frac{1}{B^2}\frac{\Delta\rho_B}{\rho_0}=\frac{\sigma_0-\sigma_B}{\sigma_B}\cdot\frac{1}{B^2}. \qquad (45.3)$$

For known values of H the resistance ρ_B in the presence of a magnetic field B may be written in the form

$$\rho_B=\rho_0(1+HB^2); \quad \sigma_B=\frac{\sigma_0}{1+HB^2}. \qquad (45.4)$$

Since $\sigma_0=e^2K_{11}$ we may, using the expressions (45.1), (45.2) and (45.3), write for H

$$H=\frac{1}{B^2}\,\frac{e^2(K_{11}-K'_{11})-\frac{e^4}{m^{*2}}\frac{(K'_{12})^2}{K'_{11}}B^2}{e^2K'_{11}+\frac{e^4}{m^{*2}}\frac{(K'_{12})^2}{K'_{11}}B^2}, \qquad (45.5)$$

or

$$H=\frac{1}{B^2}\,\frac{\langle\mu\rangle-\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle-\dfrac{\left\langle\frac{\mu^2}{1+\mu^2B^2}\right\rangle^2B^2}{\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle}}{\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle+\dfrac{\left\langle\frac{\mu^2}{1+\mu^2B^2}\right\rangle^2B^2}{\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle}}. \qquad (45.6)$$

Consider the range of field values for which we may write, using (44.4),

$$H=\frac{1}{B^2}\,\frac{\langle\mu^3\rangle B^2-B^2\dfrac{[\langle\mu^2\rangle-\langle\mu^4\rangle B^2]^2}{[\langle\mu\rangle-\langle\mu^3\rangle B^2]}}{\langle\mu\rangle-\langle\mu^3\rangle B^2+B^2\dfrac{[\langle\mu^2\rangle-\langle\mu^4\rangle B^2]^2}{[\langle\mu\rangle-\langle\mu^3\rangle B^2]}}=$$

$$=\frac{\langle\mu^3\rangle\langle\mu\rangle-\langle\mu^2\rangle^2}{\langle\mu\rangle^2}\left\{1+B^2\left[\frac{\langle\mu^4\rangle}{\langle\mu^2\rangle}-\frac{\langle\mu^2\rangle^2}{\langle\mu\rangle^2}\right]\right\}. \qquad (45.7)$$

For *weak fields* (the limit being $B=0$)

$$H=H_0=\frac{\langle\mu^3\rangle\langle\mu\rangle-\langle\mu^2\rangle^2}{\langle\mu\rangle^2}=\frac{e^2}{m^{*2}}\,\frac{\langle\tau^3\rangle\langle\tau\rangle-\langle\tau^2\rangle^2}{\langle\tau\rangle^2}=\mu_d^2\,[\mathrm{C}-\mathrm{A}^2], \qquad (45.8)$$

i.e. the magnetoresistance coefficient remains constant, and the resistance grows in proportion to B^2:

$$\rho_B=\rho_0\,(1+H_0B^2) \qquad (45.9)$$

As the field increases, the magnetoresistance coefficient (45.7) *decreases with a corresponding decrease in the rate of growth of the resistance.*

Consider *strong fields* (the limit being $B=\infty$)

$$H=\frac{1}{B^2}\,\frac{\langle\mu\rangle-\langle\mu^{-1}\rangle^{-1}}{\langle\mu^{-1}\rangle^{-1}}=\frac{1}{B^2}\,(\langle\mu\rangle\langle\mu^{-1}\rangle-1). \qquad (45.10)$$

The magnetoresistance coefficient decreases in strong magnetic fields as B^{-2}. Consequently, there is a saturation value for the resistance in strong fields:

$$\rho_B\cong\rho_0\,(1+\langle\mu\rangle\langle\mu^{-1}\rangle-1)=\rho_0\langle\mu\rangle\langle\mu^{-1}\rangle=\rho_\infty. \qquad (45.11)$$

The difference between this saturation resistance and the resistance in the absence of magnetic fields (saturation magnetoresistance)

$$\Delta\rho_\infty=\rho_\infty-\rho_0=\rho_0\,(\langle\mu\rangle\langle\mu^{-1}\rangle-1)=\rho_0\,(\langle\tau\rangle\langle\tau^{-1}\rangle-1) \qquad (45.12)$$

denends on the scattering mechanism.

If only one scattering mechanism of the type, $\tau=\tau_0 E^p$ is active,

$$\langle\tau\rangle\langle\tau^{-1}\rangle=\frac{\Gamma\left(\frac{5}{2}+p\right)\Gamma\left(\frac{5}{2}-p\right)}{\left[\Gamma\left(\frac{5}{2}\right)\right]^2}. \tag{45.13}$$

Table 13 shows the values of $[\langle\tau\rangle\langle\tau^{-1}\rangle]^{-1}$ and HB^2 for some values of p.

Table 13

p	0	$\pm 1/2$	± 1	$\pm 3/2$	± 2
$[\langle\tau\rangle\langle\tau^{-1}\rangle]^{-1}$	1	$\frac{9\pi}{32}=0.884$	$\frac{3}{5}=0.600$	$\frac{3\pi}{32}=0.295$	$\frac{3}{35}=0.086$
HB^2	0	0.116	0.400	0.705	0.914

We see that positive magnetoresistance attains a saturation value which depends on p. For $p=1/2$ the ratio $\frac{\Delta\rho_\infty}{\rho_0}$ is about 12 per cent.

In degenerate semiconductors $H=0$ in any fields. This means that in degenerate semiconductors and metals the magnetoresistance is zero, to a first approximation. In order to obtain non-zero terms one should take into account terms of higher order in the expansion (3.42).

Calculations lead to the result

$$H=\frac{H_0}{1+(\sigma RB)^2}, \tag{45.14}$$

where

$$H_0=\frac{\pi^2}{12}\left[\frac{ekT\tau(F)}{m^*F}\right]^2, \tag{45.15}$$

i.e. the magnetoresistance in weak fields is quadratic in B and reaches saturation in strong fields.

Consider a finite sample of a semiconductor with charge carriers of several types. In compliance with (43.12) and (43.15)

$$\sigma_B=\sum_\alpha e_\alpha n_\alpha\left\langle\frac{\mu_\alpha}{1+\mu_\alpha^2B^2}\right\rangle+\frac{\left(\sum_\alpha e_\alpha n_\alpha\left\langle\frac{\mu_\alpha^2}{1+\mu_\alpha^2B^2}\right\rangle\right)^2B^2}{\left(\sum_\alpha e_\alpha n_\alpha\left\langle\frac{\mu_\alpha}{1+\mu_\alpha^2B^2}\right\rangle\right)}. \tag{45.16}$$

In weak fields

$$\sigma_B=\sum_\alpha e_\alpha n_\alpha\left(\langle\mu_\alpha\rangle-\langle\mu_\alpha^3\rangle B^2\right)+$$
$$+\frac{\left[\sum\limits_\alpha e_\alpha n_\alpha\left(\langle\mu_\alpha^2\rangle-\langle\mu_\alpha^4\rangle B^2\right)\right]^2 B^2}{\left[\sum\limits_\alpha e_\alpha n_\alpha\left(\langle\mu_\alpha\rangle-\langle\mu_\alpha^3\rangle B^2\right)\right]}. \qquad (45.17)$$

Retaining only the terms of the zero and the second power of B we obtain

$$H_0=\frac{1}{B^2}\frac{\sigma_0-\sigma_B}{\sigma_B}=\frac{\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^3\rangle-\frac{\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha^2\rangle\right]^2}{\left[\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha\rangle\right]}}{\sum\limits_\alpha e_\alpha n_\alpha\langle\mu_\alpha\rangle}. \qquad (45.18)$$

Consider an intrinsic semiconductor:

$$\langle\mu_1\rangle=\mu_{nd};\quad \langle\mu_2\rangle=\mu_{pd};\quad \langle\mu_\alpha^2\rangle=\langle\mu_\alpha\rangle^2 A_\alpha;\quad A_1=A_2=A;$$
$$\langle\mu_\alpha^3\rangle=\langle\mu_\alpha\rangle^3 C_\alpha;\quad C_1=C_2=C;\quad n_1=n=n_2=p. \qquad (45.19)$$

Substituting (45.19) into (45.18) we obtain

$$H_0=\frac{C\left(\mu_p^3-\mu_n^3\right)-\frac{A^2\left(\mu_p^2-\mu_n^2\right)^2}{\mu_p-\mu_n}}{\mu_p-\mu_n}=\mu_p^2\left[C\frac{1+b^3}{1+b}-A^2(1-b)^2\right]. \qquad (45.20)$$

The meaning of the relation (45.20) is obvious. For one type of carriers, according to (45.8), $H_0=\mu_p^2[C-A^2]$. *The quantity* C *determines that part of the magnetoresistance which is due to the "screwing" of the carriers by the magnetic field; the term with* A^2 *reflects the compensating action of the Hall field* (due to finite dimensions of the sample).

For a semiconductor with unequal concentrations of electrons and holes we obtain from (45.20), denoting $\frac{n}{p}=x$,

$$H_0=C\mu_p^2\frac{(1+xb^3)}{(1+xb)}-A^2\mu_p^2\frac{(1-xb^2)^2}{(1+xb)^2}. \qquad (45.21)$$

For charge carriers of one sign

$$H_0=\mu_{p1}^2\left[C\frac{(1+xb^3)}{(1+xb)}-A^2\frac{(1+xb^2)^2}{(1+xb)^2}\right], \qquad (45.21a)$$

where μ_{p1} is the mobility of charge carriers whose concentration is p_1 (or n_1); $x=\frac{p_2}{p_1}$; $b=\frac{\mu_{p2}}{\mu_{p1}}$.

The expression for the magnetoresistance in strong fields may be obtained in the usual way from the expression (45.16).

Consider now the magnetoresistance of an infinite sample. As was already repeatedly stated, there is no Hall field in this case; the current turns through an angle φ, and there appears a non-zero current component j_z. The expression for current density takes the form of (42.2)

$$\mathbf{j}=e^2K'_{11}\mathbf{E}+\frac{e^3}{m^*}K'_{12}[\mathbf{EB}] \tag{45.22}$$

with the following "boundary" conditions: $\mathbf{E}=(E_x, 0, 0)$ for $\mathbf{B}=(0, B, 0)$. Write (45.22) for the current density components:

$$\begin{aligned} j_x&=e^2K'_{11}E_x=\sigma_{Bx}E_x; \\ j_y&=0; \\ j_z&=\frac{e^3}{m^*}K'_{12}BE_x=\sigma_{Bz}E_x. \end{aligned} \tag{45.23}$$

The conductivity σ_{Bx} determines the current in the direction of the electric field in the presence of the magnetic field. σ_{Bx} is dependent on B since K'_{11} depends on B. The quantity σ_{Bz} determines the conductivity along the z-axis in the direction of the vector $[\mathbf{EB}]$; it is termed *Hall conductivity* and *it is not equal to* σ_{Bx}. The term *magnetoresistance* refers to the *change in resistance in the direction of the field* E_x (or j_x). Find the magnetoresistance coefficient

$$H=\frac{1}{B^2}\frac{\sigma_0-\sigma_B}{\sigma_B}=\frac{1}{B^2}\frac{\langle\mu\rangle-\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle}{\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle}=\frac{1}{B^2}\frac{\left\langle\frac{\mu^3B^2}{1+\mu^2B^2}\right\rangle}{\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle}=\frac{\left\langle\frac{\mu^3}{1+\mu^2B^2}\right\rangle}{\left\langle\frac{\mu}{1+\mu^2B^2}\right\rangle} \tag{45.24}$$

In weak fields (when the terms of the order of μ^2B^2 are taken into account)

$$H=\frac{\langle\mu^3\rangle-\langle\mu^5\rangle B^2}{\langle\mu\rangle-\langle\mu^3\rangle B^2}=\frac{\langle\mu^3\rangle}{\langle\mu\rangle}\left[1+\left(\frac{\langle\mu^3\rangle}{\langle\mu\rangle}-\frac{\langle\mu^5\rangle}{\langle\mu^3\rangle}\right)B^2\right]. \tag{45.25}$$

In fields for which $\mu_d^2B^2\cong 0\,(B\longrightarrow 0)$

$$H_0=\frac{\langle\mu^3\rangle}{\langle\mu\rangle}=\mu_d^2\frac{\langle\tau^3\rangle}{\langle\tau\rangle^3}=C\mu_d^2. \tag{45.26}$$

Comparing the expression (45.26) with (45.8) we see that the respective H_0s are different because of a factor in front of μ_d^2 which is equal to C for an infinite sample and to C—A², for a finite sample. In other words, *in a finite sample the magnetoresistance is less pronounced since it is compensated by the Hall field.* It follows from (45.26) that

$$\mu_d = \left(\frac{H_0}{C}\right)^{\frac{1}{2}}. \tag{45.27}$$

In a strong magnetic field

$$H_\infty = \frac{\langle\mu\rangle}{\langle\mu^{-1}\rangle} = \frac{e^2\langle\tau\rangle^2}{m^{*2}} \frac{1}{\langle\tau\rangle\langle\tau^{-1}\rangle} = \mu_d^2 \frac{1}{\langle\tau\rangle\langle\tau^{-1}\rangle}. \tag{45.28}$$

At the same time in a finite sample H_∞ is determined by the relation (45.10). We see from here that *in an infinite sample H is independent of B in weak as well as in strong fields*. This means that $\rho_B = \rho_0(1 + HB^2)$ for $B \to 0$ and $B \to \infty$. The last conclusion is understandable since as the field B increases the charge carriers rotate in circles of ever decreasing radius $r = \frac{p}{eB}$ (p is the momentum) and drift in the direction of the z-axis with the velocity $u = \frac{B}{E}$ with the result that the *resistance increases to infinity*. Table 14 shows the values of $\frac{H}{\mu_d^2}$ in an infinite and in a finite sample.

It follows from Table 14 that the magnetoresistance coefficient in an infinite semiconductor sample in a weak field is several times (depending on p) greater than that of a finite sample. For instance, for $p = 1/2$ it is 11 times greater. This difference grows with the field intensity. Figure 68 shows the dependence of magnetoresistance of InSb on the shape of the sample. The figure shows that *the part played by the compensating Hall field is the greater the smaller is the width-to-length ratio of the sample, the*

Table 14

p	0	$-1/2$	$1/2$	$+1$	$3/2$	2	$5/2$
$\frac{H}{\mu_d^2}$ (fin)	0	0.38	0.11	0.56	2.98	7.7	9.2
$\frac{H}{\mu_d^2}$ (infin)	1	$\frac{9\pi}{16}=1.77$	$\frac{27}{64}=1.33$	$\frac{63}{25}=2.52$	$\frac{15\pi}{8}=5.90$	15.7	$\frac{945\pi}{64}=46.4$
$\frac{\langle\tau^2\rangle^2}{\langle\tau\rangle^4}$	1	$\frac{9\pi^2}{64}=1.39$	$\frac{9\pi^2}{64}\cdot\frac{15}{16}=$ $=1.22$	$\frac{49}{25}=1.96$	3.72	8	37.2

disc-shaped sample (Corbino disc) being equivalent to an "infinite" sample.

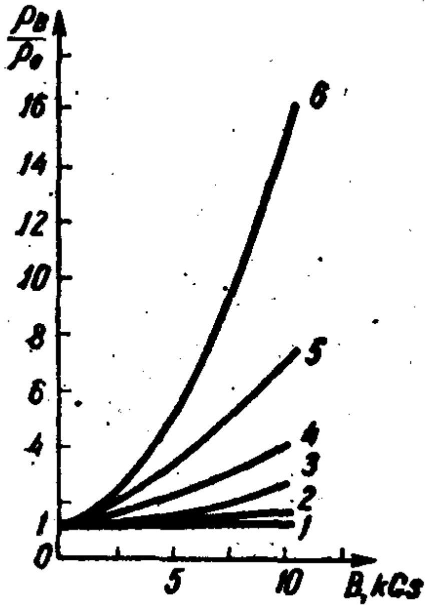


Fig. 68. The dependence of the magnetoresistive effect on the sample length-to-width ratio:

1 – ∞; 2 – 5; 3 – 2; 4 – 1; 5 – 0.4; 6 – 0 (Corbino disc)

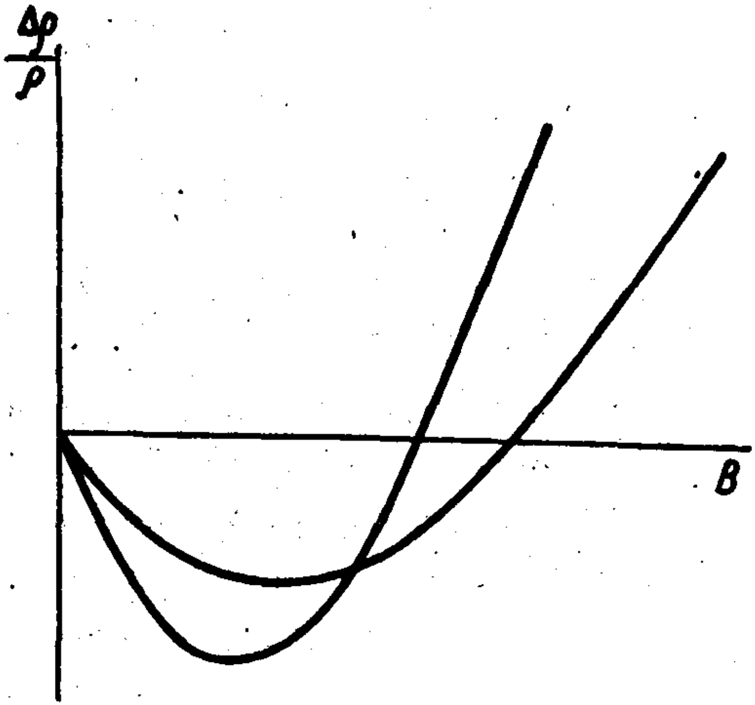


Fig. 69. Negative magnetoresistance in quicksilver telluride

In conclusion of this section discuss the magnetoresistance in an infinite sample for the case of charge carriers of several types. The expression for the current density is

$$\mathbf{j}=\left(\sum_{\alpha} e_{\alpha}^{2}K'_{11\,(\alpha)}\right)\mathbf{E}+\left(\sum_{\alpha}\frac{e_{\alpha}^{3}}{m_{\alpha}^{*}}K'_{12\,(\alpha)}\right)[\mathbf{EB}], \qquad (45.29)$$

or

$$j_x=\left(\sum_{\alpha} e_{\alpha}^{2}K'_{11\,(\alpha)}\right)E_x=\sigma_{Bx}E_x, \qquad (45.30)$$

$$j_y=0; \qquad (45.31)$$

$$j_z=\sum_{\alpha}\frac{e_{\alpha}^{3}}{m_{\alpha}^{*}}K'_{12}BE_x=\sigma_{Bz}E_x. \qquad (45.32)$$

The magnetoresistance coefficient

$$H=\frac{1}{B^2}\cdot\frac{\sum\limits_{\alpha} e_{\alpha}n_{\alpha}\left[\langle\mu_{\alpha}\rangle-\left\langle\frac{\mu_{\alpha}}{1+\mu_{\alpha}^{2}B^{2}}\right\rangle\right]}{\sum\limits_{\alpha} e_{\alpha}n_{\alpha}\left\langle\frac{\mu_{\alpha}}{1+\mu_{\alpha}^{2}B^{2}}\right\rangle}. \qquad (45.33)$$

Its value in weak fields is

$$H = H_0 = \frac{\sum\limits_{\alpha} e_\alpha n_\alpha \langle \mu_\alpha^2 \rangle}{\sum\limits_{\alpha} e_\alpha n_\alpha \langle \mu_\alpha \rangle} = \frac{\sum\limits_{\alpha} \sigma_\alpha H_{0\alpha}}{\sum\limits_{\alpha} \sigma_\alpha}; \qquad (45.34)$$

and in strong fields

$$H = H_\infty = \frac{\sum\limits_{\alpha} e_\alpha n_\alpha \langle \mu_\alpha \rangle}{\sum\limits_{\alpha} e_\alpha n_\alpha \langle \mu_\alpha^{-1} \rangle} = \frac{\sum\limits_{\alpha} \sigma_\alpha}{\sum\limits_{\alpha} \frac{\sigma_\alpha}{H_{\infty\alpha}}}; \qquad H^{-1} = \frac{\sum\limits_{\alpha} \sigma_\alpha H_{\infty\alpha}^{-1}}{\sum\limits_{\alpha} \sigma_\alpha}. \qquad (45.35)$$

Consider a semiconductor in the mixed conductivity range (for a weak field):

$$H_0 = \frac{\mathrm{C}_p p \mu_p^3 - \mathrm{C}_n n \mu_n^3}{p\mu_p - n\mu_n} = \frac{\mathrm{C} p \mu_p^3}{p\mu_p} \frac{1 + xb^3}{1 + xb} = H_p \frac{1 + xb^3}{1 + xb}, \qquad (45.36)$$

where

$$H_p = \mu_p^2 \mathrm{C}_p = \mu_p^2 \frac{\langle \tau_p^3 \rangle}{\langle \tau_p \rangle^3}. \qquad (45.37)$$

A similar relationship may be obtained for the case of a strong field as well.

The theory of magnetoresistance discussed above is applicable to a semiconductor of the simplest kind with a scalar effective mass. In semiconductors with an intricate energy-band pattern all sorts of anomalous phenomena may take place, such as, for instance, negative magnetoresistance, anisotropy of magnetoresistance, etc.

Figure 69 shows the magnetic field dependence of the magnetoresistance of cadmium telluride-quicksilver telluride solid solution $\mathrm{Cd}_x\mathrm{Hg}_{1-x}\mathrm{Te}$ samples for $x \cong 0.25$. The negative magnetoresistance area is well visible. Temperature affects magnetoresistance since it affects the concentration of charge carriers. Figure 70 shows, by way of an example, the temperature dependence of the magnetoresistance of a cadmium telluride-quicksilver telluride solid solution sample. A maximum is visible on the $\frac{\Delta\rho}{\rho_0}(T)$ curve. Figure 71 shows the temperature dependence of the Hall coefficient for the same sample. Comparison of the figures shows that the magnetoresistance maximum is within the Hall coefficient inversion area, where the change from one type of charge carrier dominant in galvanomagnetic effects to another takes place. The decrease of magnetoresistance observed at room temperature is due to the fact that the sample becomes almost intrinsic with the electron gas in a degenerate state.

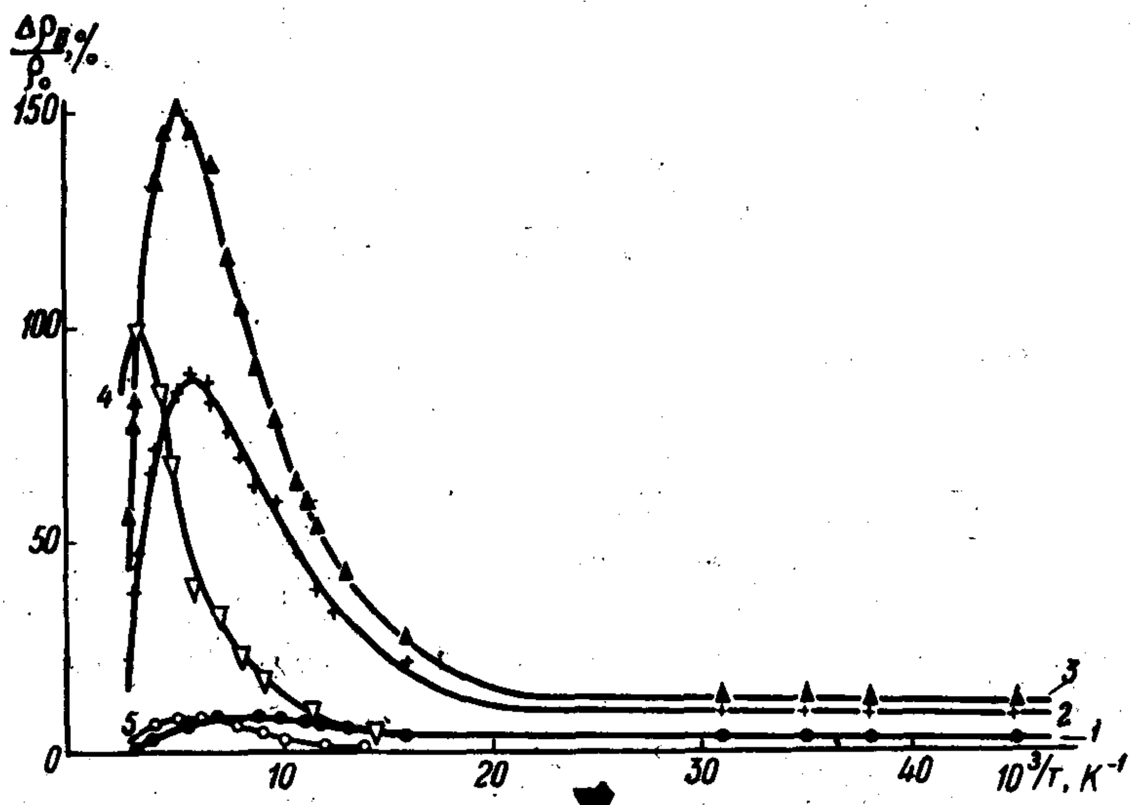


Fig 70. The temperature dependence of magnetoresistance in the cadmium and quicksilver tellurides solid solution Cd_xHg_{1-x} ($x \cong 0.25$)

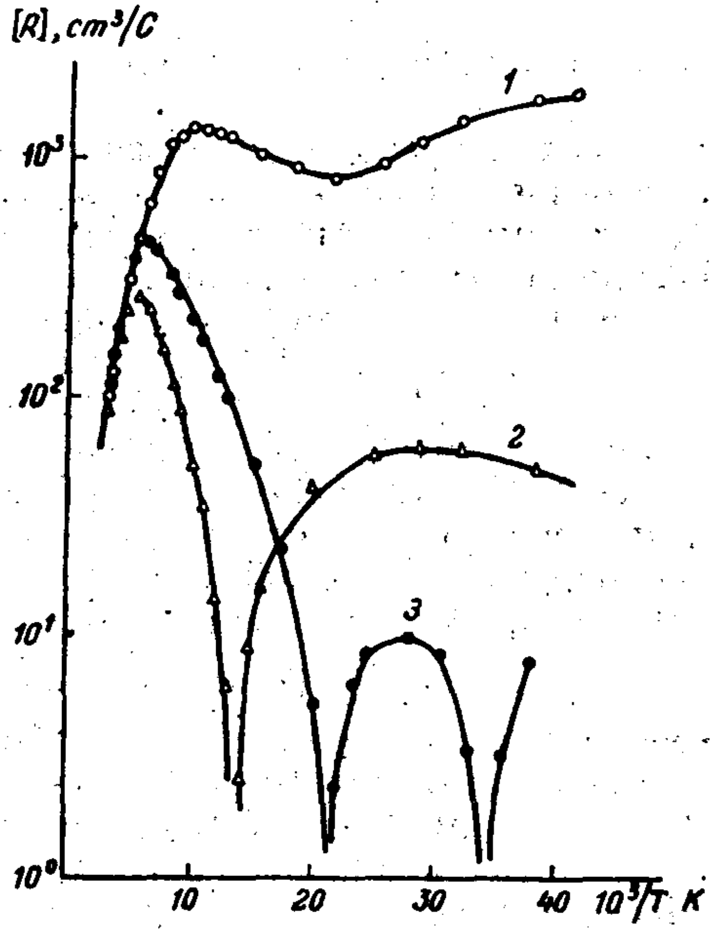


Fig. 71. The temperature dependence of the Hall coefficient in a Cd_xHg_{1-x} Te ($x \cong 0.25$) sample:

1—2.68; 2—10.78; 3—19.2 kGs

Summary of Sec. 45

1. The change in the resistance of a semiconductor due to magnetic fields is caused by the "screwing" action exercised by these fields on the charge carriers. The magnetoresistance effect is described by the coefficient H:

$$H=\frac{1}{B^2}\frac{\rho_B-\rho_0}{\rho_0}=\frac{1}{B^2}\frac{\rho_0-\rho_B}{\rho_{B_0}}, \qquad (45.1s)$$

which enables the resistance ρ_B to be written in the form

$$\rho_B=\rho_0(1+HB^2). \qquad (45.2s)$$

2. In a finite semiconductor sample in weak fields H remains constant. In strong fields $H \sim B^{-2}$, and this results in the saturation of the resistance.

3. In an infinite semiconductor sample H is constant both in weak and in strong fields, but its values are different.

46. HEAT CONDUCTIVITY OF SEMICONDUCTORS

If a temperature gradient ∇T is built up in a substance, the result will be an energy flux **W** *flowing in the direction opposite to ∇T:*

$$\mathbf{W}=-\varkappa\nabla T \qquad (46.1)$$

$\varkappa=\frac{|\mathbf{W}|}{|\nabla T|}$ *is termed heat conductivity.* Numerically, it is equal to the amount of energy passing per unit time through a unit cross section of a sample between the ends of which a temperature difference of one degree is maintained. In the SI system $\varkappa$ is measured in J/(s·m²·K/m) = W/(m·K). This definition of **W** presumes all other processes in the substance to be inoperative. Because of heat conductivity heat is transferred from the "heater" to the "cooler". In a system which does not include a heat source and a heat sink, but where a temperature gradient has been built up, heat conductivity results in the levelling out of the temperature, i.e. in thermal equilibrium.

Two different mechanisms are responsible for heat conduction. *Heat conductivity due to the charge carrier motion is termed electron or hole heat conductivity.* It is described by means of the heat conductivity $\varkappa_e$. The second mechanism is connected with the lattice vibrations. The lattice atoms (or ions) oscillating around their respective equilibrium positions exchange energy with each other. When a temperature gradient has been built up in a substance, this energy exchange proceeds in such a manner that energy is transmitted from an atom which oscillates more intensely to an atom which oscillates less intensely, i. e. in the direction of lower temperatures. *Heat conductivity due to lattice vibrations is termed lattice, or phonon, heat conductivity; it is described by the quantity* $\varkappa_L$. Full heat conductivity may thus be described by the quantity $\varkappa$:

$$\varkappa = \varkappa_e + \varkappa_L. \tag{46.2}$$

The value of $\varkappa_L$ should be related to the elastic properties of the solid, $\varkappa_e$—to the charge carrier concentration. In the dielectrics $\varkappa_L \gg \varkappa_e$. In metals the opposite may be the case: $\varkappa_e \gg \varkappa_L$. In semiconductors the value of $\varkappa_e$ should strongly depend on their composition and on temperature. We will consider below only the heat conductivity due to the charge carriers. To describe it write the expressions (38.20) for the current density and (38.21) for the energy flux density presuming the magnetic field to be absent ($\mathbf{B}=0$):

$$\mathbf{j} = \left(e\mathbf{E} - T\nabla \frac{F}{T}\right) eK_{11} - eK_{21}\frac{\nabla T}{T} \tag{46.3}$$

and

$$\mathbf{W} = \left(e\mathbf{E} + T\nabla \frac{F}{T}\right) K_{21} - K_{31}\frac{\nabla T}{T}. \tag{46.4}$$

Find the expression for the conductivity $\varkappa_e$ which is determined by the condition

$$\mathbf{W} = -\varkappa_e \nabla T, \tag{46.5}$$

in the assumption

$$\mathbf{j} = 0. \tag{46.6}$$

Determine the quantity $\left(e\mathbf{E} - T\nabla \frac{F}{T}\right)$ from the condition (46.6) and the expression (46.3), and substitute it into (46.4):

$$e\mathbf{E} - T\nabla \frac{F}{T} = \frac{K_{21}}{K_{11}T}\nabla T. \tag{46.7}$$

$$\mathbf{W} = \frac{K_{21}^2}{K_{11}T}\nabla T - K_{31}\frac{\nabla T}{T} = -\frac{K_{31}K_{11} - K_{21}^2}{K_{11}T}\nabla T. \tag{46.8}$$

It follows from (46.5) and (46.8) that

$$\varkappa_e = \frac{K_{31}K_{11} - K_{21}^2}{K_{11}T}, \tag{46.9}$$

or

$$\varkappa_e = \frac{n}{m^*}\frac{\langle E^2\tau\rangle\langle E\tau\rangle - \langle E^2\tau\rangle^2}{\langle E\tau\rangle T}. \tag{46.10}$$

In the assumption that the relaxation time is a power function of energy we obtain, according to (39.30), for a *non-degenerate semi-*

conductor

$$\varkappa_e=\frac{n}{m^*}\frac{5k^2T\tau_0'}{4}\frac{\left\{7\frac{\Gamma\left(\frac{9}{2}+p\right)\Gamma\left(\frac{5}{2}+p\right)}{\Gamma\left(\frac{9}{2}\right)\Gamma\left(\frac{5}{2}\right)}-5\frac{\Gamma^2\left(\frac{7}{2}+p\right)}{\Gamma^2\left(\frac{7}{2}\right)}\right\}}{\frac{\Gamma\left(\frac{5}{2}+p\right)}{\Gamma\left(\frac{5}{2}\right)}}. \qquad (46.11)$$

The expression for $\varkappa_e$ contains an unknown constant τ_0'. It may be cancelled out if the ratio $\varkappa_e/\sigma$ is considered. To be precise, consider the *ratio* $\frac{\varkappa_e}{\sigma T}=L$. *It is independent of* τ_0' *and is termed the Lorentz number:*

$$L=\frac{\varkappa_e}{\sigma T}=\frac{K_{31}K_{11}-K_{21}^2}{e^2K_{11}^2T^2}. \qquad (46.12)$$

For the Lorentz number, according to (46.11), we have the expression (for the case of a non-degenerate semiconductor):

$$L=\frac{K_{31}K_{11}-K_{21}^2}{e^2K_{11}^2T^2}=\frac{k^2}{e^2}\left(\frac{5}{2}+p\right). \qquad (46.13)$$

We see that the *Lorentz number depends only on p*. Expressing $\varkappa_e$ in terms of σ and L we obtain

$$\frac{\varkappa_e}{\sigma}=LT=\frac{k^2T}{e^2}\left(\frac{5}{2}+p\right), \qquad (46.14)$$

$$\varkappa_e=\sigma LT=\frac{k^2}{e^2}\left(\frac{5}{2}+p\right)\sigma T. \qquad (46.15)$$

The ratio (46.14) *expresses the Wiedemann-Franz law. For a degenerate semiconductor the Lorentz number is*

$$L=\frac{\varkappa_e}{\sigma T}=\frac{\pi^2}{3}\frac{k^2}{e^2}. \qquad (46.16)$$

To obtain this result higher terms of the expansion (3.42) should be invoked, i.e. this is the second approximation.

If a *substance contains charge carriers of several types* the expression for $\varkappa_e$ may be obtained from the equations

$$\mathbf{j}=\sum_\alpha\left(e_\alpha^2K_{11(\alpha)}\mathbf{E}-T\nabla\frac{F_\alpha}{T}e_\alpha K_{11(\alpha)}\right)-\sum_\alpha e_\alpha K_{21(\alpha)}\frac{\nabla T}{T}=$$
$$=\left(\sum_\alpha e_\alpha^2K_{11(\alpha)}\right)\mathbf{E}-\left(\sum_\alpha e_\alpha K_{11(\alpha)}\nabla F_\alpha\right)+$$
$$+\left[\sum_\alpha e_\alpha\left(K_{11(\alpha)}F_\alpha-K_{21(\alpha)}\right)\right]\frac{\nabla T}{T}=0 \qquad (46.17)$$

and

$$\mathbf{W}=\Big(\sum_{\alpha} e_{\alpha}^{2}K_{21(\alpha)}\Big)\mathbf{E}-\Big(\sum_{\alpha} K_{21(\alpha)}\nabla F_{\alpha}\Big)+ \\ +\Big[\sum_{\alpha}(K_{21(\alpha)}E_{\alpha}-K_{31(\alpha)})\Big]\frac{\nabla T}{T}. \qquad (46.18)$$

In the expressions (46.17) and (46.18) the Fermi level has a subscript α. It is important because when calculating the kinetic coefficients we took E_0 (E_c or E_v) for the energy scale origin, i.e. *the origin of the energy scale is different for different types of charge carriers, and therefore the value of the Fermi level for each type of charge carriers should be expressed in the corresponding reference system.* For this reason the expressions for F for the electrons and holes are different. Their sum

$$F_n+F_p=(F-E_c)+(E_v-F)=-\Delta E_0, \qquad (46.19)$$

where F is the "true" Fermi level position in some energy reference system common to both the electrons and holes.

For different F_{α} ∇F_{α} may be equal in absolute value, since they determine the force acting on the charge carriers of the type α and resulting from a non-uniform impurity distribution. Put

$$\nabla F_{\alpha}=-e_{\alpha}\mathbf{E}^{i}. \qquad (46.20)$$

Express the quantity $\mathbf{E}+\mathbf{E}^{i}$ from (46.17):

$$\mathbf{E}+\mathbf{E}^{i}=\frac{\sum_{\alpha}(K_{21(\alpha)}-F_{\alpha}K_{11(\alpha)})e_{\alpha}}{\Big(\sum_{\alpha}e_{\alpha}^{2}K_{11(\alpha)}\Big)}\frac{\nabla T}{T} \qquad (46.21)$$

ınd substitute into (46.18) to obtain

$$\mathbf{W}=-\varkappa_{e}\nabla T=\left\{\Big(\sum_{\alpha}e_{\alpha}K_{21(\alpha)}\Big)\frac{\Big[\sum_{\alpha}e_{\alpha}(K_{21(\alpha)}-F_{\alpha}K_{11(\alpha)})\Big]}{\Big(\sum_{\alpha}e_{\alpha}^{2}K_{11(\alpha)}\Big)}+\right. \\ \left.+\sum_{\alpha}(F_{\alpha}K_{21(\alpha)}-K_{31(\alpha)})\right\}\frac{\nabla T}{T}. \qquad (46.22)$$

Consider a semiconductor of mixed conductivity. In compliance with (46.22) it may be written

$$-\varkappa_e = \frac{(K_{21p}-K_{21n})\,[K_{21p}-F_pK_{11p}-K_{21n}+F_nK_{11n}]}{(K_{11p}+K_{11n})\,T} + \frac{F_pK_{21p}-K_{31p}+F_nK_{21n}-K_{31n}}{T} =$$

$$= \frac{K_{21p}^2\left(1-\frac{K_{21n}}{K_{21p}}\right)^2 - K_{11p}\left(F_p-F_n\frac{K_{11n}}{K_{11p}}\right)\left(1-\frac{K_{21n}}{K_{21p}}\right)K_{21p}}{K_{11p}\left(1+\frac{K_{11n}}{K_{11p}}\right)T} + \frac{K_{21p}}{T}\left(F_g+F_n\frac{K_{21n}}{K_{21p}}\right) - \frac{K_{31p}}{T}\left(1+\frac{K_{31n}}{K_{31p}}\right). \qquad (46.23)$$

The conductivity may be represented in the form

$$\sigma = e_p^2\,(K_{11p}+K_{11n}) = e^2K_{11p}\left(1+\frac{K_{11n}}{K_{11p}}\right). \qquad (46.24)$$

Find the value of the Lorentz number, or, to be more precise, of Le^2T^2:

$$Le^2T^2 = -\frac{K_{21p}^2}{K_{11p}^2}\,\frac{\left(1-\frac{K_{21n}}{K_{21p}}\right)^2}{\left(1+\frac{K_{11n}}{K_{11p}}\right)^2} + \frac{K_{21p}\left(1-\frac{K_{21n}}{K_{21p}}\right)\left(F_p-F_n\frac{K_{11n}}{K_{11p}}\right)}{K_{11p}\left(1+\frac{K_{11n}}{K_{11p}}\right)^2} - \frac{K_{21p}}{K_{11p}}\,\frac{\left(F_p+F_n\frac{K_{21n}}{K_{21p}}\right)}{\left(1+\frac{K_{11n}}{K_{11p}}\right)} + \frac{K_{31p}}{K_{11p}}\,\frac{\left(1+\frac{K_{31n}}{K_{31p}}\right)}{\left(1+\frac{K_{11n}}{K_{11p}}\right)}. \qquad (46.25)$$

Consider the ratios of the kinetic coefficients

$$\frac{K_{11n}}{K_{11p}} = xb = -\frac{n\mu_n}{p\mu_p}; \qquad (46.26)$$

$$\frac{K_{21n}}{K_{21p}} = xb\,\frac{\langle\tau_p\rangle\,\langle\langle\tau_n\rangle\rangle}{\langle\langle\tau_p\rangle\rangle\,\langle\tau_n\rangle}; \qquad (46.27)$$

$$\frac{K_{31n}}{K_{31p}} = xb\,\frac{\langle\langle\langle\tau_n\rangle\rangle\rangle\,\langle\tau_p\rangle}{\langle\tau_n\rangle\,\langle\langle\langle\tau_p\rangle\rangle\rangle}; \qquad (46.28)$$

$$\frac{K_{21p}}{K_{11p}} = \frac{\langle\langle\tau_p\rangle\rangle}{\langle\tau_p\rangle};\quad \frac{K_{31p}}{K_{11p}} = \frac{\langle\langle\langle\tau_p\rangle\rangle\rangle}{\langle\tau_p\rangle}. \qquad (46.29)$$

The kinetic coefficient ratios for the case of a single-mechanism scattering may be expressed in terms of the Γ-function:

$$\frac{K_{21p}}{K_{11p}}=kT\frac{\Gamma\left(\frac{7}{2}+p\right)}{\Gamma\left(\frac{5}{2}+p\right)}=\left(\frac{5}{2}+p\right)kT; \qquad (46.30)$$

$$\frac{K_{31p}}{K_{11p}}=(kT)^2\frac{\Gamma\left(\frac{9}{2}+p\right)}{\Gamma\left(\frac{5}{2}+p\right)}=\left(\frac{5}{2}+p\right)\left(\frac{7}{2}+p\right)(kT)^2. \qquad (46.31)$$

From the above expressions (46.26-29) it follows also that

$$\frac{K_{31n}}{K_{31p}}=xb\frac{\Gamma\left(\frac{9}{2}+p_n\right)}{\Gamma\left(\frac{9}{2}+p_p\right)}; \quad \frac{K_{21n}}{K_{21p}}=xb\frac{\Gamma\left(\frac{7}{2}+p_n\right)}{\Gamma\left(\frac{7}{2}+p_p\right)}, \qquad (46.32)$$

where p_n and p_p determine the energy dependence of electron and hole relaxation time.

The kinetic coefficients' ratio $K_{n1(\alpha)}$ is equal to the ratio of concentrations and mobilities (i.e. of conductivities), therefore

$$\frac{K_{r1(n)}}{K_{r1(p)}}=xb\frac{\Gamma\left(\frac{3}{2}+r+p_n\right)}{\Gamma\left(\frac{3}{2}+r+p_p\right)}=\frac{\sigma_n}{\sigma_p}\frac{\Gamma\left(\frac{3}{2}+r+p_n\right)}{\Gamma\left(\frac{3}{2}+r+p_p\right)}. \qquad (46.33)$$

Taking into account the expressions for kinetic coefficient ratios find L (for $p_n=p_p$):

$$Le^2T^2=-\left(\frac{5}{2}+p\right)^2(kT)^2\frac{\left(1-\frac{\sigma_n}{\sigma_p}\right)^2}{\left(1+\frac{\sigma_n}{\sigma_p}\right)^2}+$$

$$+\left(\frac{5}{2}+p\right)(kT)\frac{\left(1-\frac{\sigma_n}{\sigma_p}\right)\left(F_p-F_n\frac{\sigma_n}{\sigma_p}\right)}{\left(1+\frac{\sigma_n}{\sigma_p}\right)^2}-$$

$$-\left(\frac{5}{2}+p\right)(kT)\frac{\left(F_p+F_n\frac{\sigma_n}{\sigma_p}\right)}{\left(1+\frac{\sigma_n}{\sigma_p}\right)}+\left(\frac{5}{2}+p\right)\left(\frac{7}{2}+p\right)(kT)^2. \qquad (46.34)$$

Consider the case $\sigma_h/\sigma_p=1$. From (46.34) we write

$$Le^2T^2=\left(\frac{5}{2}+p\right)\left[\left(\frac{7}{2}+p\right)+\frac{\Delta E_g}{2kT}\right](kT)^2 \qquad (46.35)$$

and

$$L=\frac{k^2}{e^2}\left(\frac{5}{2}+p\right)\left[\left(\frac{7}{2}+p\right)+\frac{\Delta E_0}{2kT}\right]. \qquad (46.36)$$

Comparing (46.36) with (46.13) we see that for *two types of charge carriers the Lorentz coefficient and*, consequently, *the heat conductivity*, as well, *increases* $\left[\frac{7}{5}+p+\frac{\Delta E_0}{2kT}\right]$ *times as compared to the case of one type of charge carriers.*

There is a very clear physical explanation for this result. As a temperature gradient is built up, a diffusion current of charge carriers begins to flow with the results that charge carriers of different types become separated, and an electric field appears which hinders the motion of the charge carriers leading to this separation.

The initial condition (46.6) requires that $\mathbf{j}=0$. Consequently, *the carrier fluxes in the direction of* ∇T *and in the opposite direction should be equal*; however, since the energy of the charge carriers moving along ∇T (to the hot end) is less than that of the charge carriers moving against ∇T (to the cold end), *energy is transported from the hot to the cold end of the sample without the charge being transported, as well.* The heat conductivity is in such cases rather small and independent of the Fermi level, since the entire effect is based on the difference of charge carrier energy fluxes flowing from the hot and the cold ends of the sample. If, however, the semiconductor contains carriers of *two types*, their diffusion currents will result in oppositely directed electric fields. Because of this the *resultant electric field will be small* (it must satisfy the condition $\mathbf{j}=0$) *and will not hinder the movement of the charge carriers, and the particle currents will be high.* The combined carrier current is caused by the concentration gradient. It will, however, not lead to a perfect levelling of concentrations since *charge carrier pairs are continuously generated at the hot end, while at the cold end charge carrier recombination prevails over generation.* Each act of charge carrier pair recombination is accompanied by the release of an amount ΔE_0 of energy which was spent to generate it at the hot end. In other words, *each charge carrier pair transports additional energy* ΔE_0, *and this is the cause of such a drastic increase in the value of the Lorentz number and, as a consequence of the heat conductivity.*

For an intrinsic semiconductor $\frac{\sigma_n}{\sigma_p}=b$ and

$$L=\frac{k^2}{e^2}\left(\frac{5}{2}+p\right)\left\{\left(\frac{7}{2}+p\right)+\frac{2b\Delta E_0}{(1+b)^2kT}-\left(\frac{5}{2}+p\right)\left(\frac{1-b}{1+b}\right)^2\right\}. \qquad (46.37)$$

It follows from (46.37) that for $b=1$ the expression for L coincides with (46.36), since in this case $\sigma_n=\sigma_p$ For $b\neq 1$ L decreases as compared to the case of $b=1$ (or $\sigma_n=\sigma_p$), this decrease being the greater the greater is the difference in mobilities.

For $b\gg 1$ the expression (46.37) may be presented in the form

$$L=\left(\frac{5}{2}+p\right)\left[1+\frac{2\Delta E_0}{bkT}\right]\frac{k^2}{e^2}. \qquad (46{:}38)$$

For $b\to\infty$ the expression (46.38) reduces to (46.13) for this is tantamount to single carrier type conductivity.

'47. THERMOELECTRIC PHENOMENA

Thermoelectric phenomena consist of three effects: the Seebeck effect, the Peltier effect and the Thomson effect. Let us discuss these effects from a qualitative standpoint.

1. **Seebeck, or thermoelectric, effect.** Suppose we have two samples, *1* and *2*, of different substances in contact with each other (Fig. 72). *If the temperature of the two contacts is different, i.e. $T+dT$ and T, the closed circuit will carry a current termed thermoelectric.* If the circuit is cut at an arbitrary point, *a potential difference will appear at the terminals termed thermoelectric electromotive force* (t. e. m. f.). Seebeck, who discovered this phenomenon, found that the potential difference $d\mathcal{E}_{12}$ in an open circuit depends on temperature difference and on the nature of the substances:

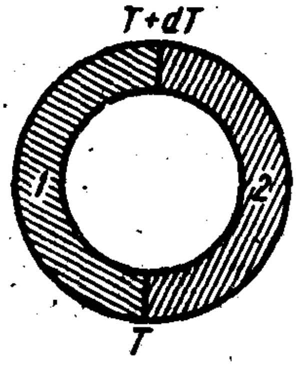


Fig. 72. Schematic drawing of the method of observing the thermoelectric effect

$$d\mathcal{E}_{12}=\alpha_{12}dT. \qquad (47.1)$$

$d\mathcal{E}_{12}$ (and α_{12}) *is assumed to be positive if the potential of the "hot" contact exceeds that of the "cold" contact*, as in Fig. 72. Correspondingly, $d\mathcal{E}_{12}=-d\mathcal{E}_{21}$, or $\alpha_{12}=-\alpha_{21}$. It follows from Fig. 72 that *if $d\mathcal{E}_{12}>0$ the current will flow clockwise, if $d\mathcal{E}_{12}<0$ the current will flow counterclockwise.* In other words, $d\mathcal{E}_{12}>0$ means that a positive charge goes over from *1* to *2* in the "hot" contact, and $d\mathcal{E}_{12}<0$—in the "cold" contact. The quantity α_{12} is characteristic of a pair of substances; it is termed *differential t.e.m.f.* The quantity α_{12} may depend on temperature T, therefore t.e.m.f. in a circuit with a finite temperature difference T_2-T_1 of its contacts is equal to

$$\mathcal{E}_{12}=\int_{T_1}^{T_2}\alpha_{12}(T)\,dT. \qquad (47.2)$$

Let us see how the t.e.m.f. at the point of contact of two metals is established. As is well known, the position of the electrochemical potential (Fermi level) in any system in thermodynamic equilibrium is everywhere the same. The Fermi levels of two metals brought into contact should, therefore, coincide. But if the electron concentrations in the two metals are different, their Fermi energies measured from the bottom of the conduction band of each metal will be different, and, hence, E_{c1} will not coincide with E_{c2}. *The difference $E_{c2}-E_{c1}$ constitutes a potential barrier at the contact; it came to be known as the internal contact potential difference* U^i *(internal work function).* U^i *is determined by the difference of the Fermi energies of the initial metals:*

$$U^i=\frac{F_1-F_2}{e}. \tag{47.3}$$

Hence, an electric field at the contact is localized in a thin layer adjoining the contact.

If a closed circuit is assembled from the two metals, U^i will be established at both contacts.

The electric field in both contacts will, obviously, be directed similarly: from greater to lower F. This means that if we move along a closed path around the circuit, in one contact the direction of motion will coincide with the field, while in another it will be opposite to it. Hence, *the circulation of the vector* **E** *will be zero.*

Suppose now that the temperature of one contact was changed by the amount dT. U^i will change, too, since the Fermi energy depends on temperature.

But if internal contact potential difference has changed, so, too, has the electric field in one of the contacts, and for this reason the circulation of the vector **E** will no longer be zero, i.e. *an e.m.f. will be established in the closed circuit.*

2. **Peltier phenomenon, or electrothermal Peltier effect.** *When electric current flows through a contact between two substances, heat is released or absorbed in the contact depending upon the direction of the current.* This phenomenon was named after Peltier. Should the direction of the current change, the effect will change sign, too. The amount of heat released and its sign depend on the nature of contacting substances, on current intensity and on the time it has been flowing, i.e. the amount of heat released is proportional to the charge $dq=I\,dt$ that has passed through the contact:

$$dQ_{12}=\Pi_{12}I\,dt, \tag{47.4}$$

dQ_{12} (and Π_{12}) are taken to mean that the current flows from the first substance to the second, while dQ_{21} (and Π_{21}) refer to the opposite case (current flows from the second substance to the first).

Evidently,

$$dQ_{21} = \Pi_{21} I\, dt = -dQ_{12} = -\Pi_{12} I\, dt. \qquad (47.5)$$

Heat released is assumed to be positive. Hence, $\Pi_{ij} > 0$ if the current from the sample i to j is accompanied by heat release.

It was discovered that *if the direction of the external current and that of the thermal current which sets in when a certain contact is heated coincide, this contact is cooled.* This may be easily understood on the basis of the energy conservation law. If we heat some contact with the result that a thermal current sets in, its direction should be such that the heat delivered to the contact should be absorbed. Therefore, if the external current is of the same direction as the thermal one, the contact should cool.

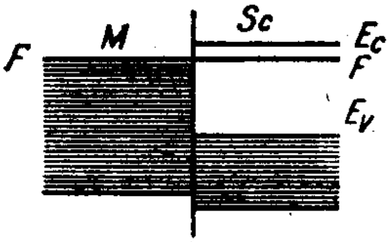


Fig. 73. Energy band pattern of the contact of a metal and an electron-type semiconductor (with equal electron work functions)

The reason for the Peltier phenomenon may be readily understood if one takes a look at Fig. 73 which depicts the energy band pattern of a metal-semiconductor contact. As we know, in a metal the conductivity is realized by electrons whose energy is close to the Fermi level. In a semiconductor (in this case, electron-type) the current is carried by the electrons of the conduction band. Figure 73 shows that the average energy of conduction electrons in the semiconductor exceeds that in the metal by an amount no less than $E_c - F$. To go over from the metal to the semiconductor the electrons have to negotiate a potential barrier at least $E_c - F$ high, and in order to do this they must receive energy from the lattice. This will cause the metal in the region of contact to cool. Evidently, the energy required will depend on the number of the electrons that have passed through the contact, i.e. on the transported charge. If the current is reversed, the electrons entering the metal will have excess energy as compared with conduction electrons of the metal. The former will establish thermodynamical equilibrium with the latter by surrendering the excess energy (no less than $E_c - F$) to the lattice with the result that heat will be released at the contact. It follows from the Peltier effect mechanism that the values of the Peltier coefficients for metal-metal contacts should be much smaller than those for metal-semiconductor or semiconductor-semiconductor contacts.

The Peltier phenomenon may be approached from a somewhat different standpoint. There is an internal contact field across the contact of two substances. When current flows across the contact the field either promotes it, or hinders it. *If the current flows against this internal contact field, the external source will have to spend additional energy which will be released in the contact and* *will heat it. If, on the contrary, the current flows along the contact field, it may be supported by this field, the field doing the work or transporting the charges. The energy needed for this purpose is taken from the substance which thereby is cooled in the region of the contact.*

The Seebeck and Peltier effects are not only contact effects but volume effects, as well. *They may be observed in the volume of a non-uniform semiconductor.*

3. Thomson phenomenon, or electrothermal Thomson effect. In a semiconductor with a non-uniform temperature distribution carrier concentration will be higher in regions of elevated temperatures, therefore a temperature gradient leads to a carrier concentration gradient resulting in carrier diffusion current. This, in turn, leads to electroneutrality being disturbed. The separation of charges results in an electric field which hinders this separation. Hence, *if there is a temperature gradient in a semiconductor, there will also be a volume electric field* $\mathbf{E}^i$.

Suppose now that under the action of an external electric field $\mathbf{E}$ electric current is flowing through the sample. If the current flows against the field $\mathbf{E}^i$, the external field will have to perform work to transport the charges in relation to the field $\mathbf{E}^i$, and this will lead to heat being released, i.e. additional Joule heat will be dissipated. If the current (or the external field $\mathbf{E}$) is directed along $\mathbf{E}^i$, $\mathbf{E}^i$ will itself do the work of transporting the charge carriers to promote the current. In this case the external source will have to spend less energy to support the current than in the absence of the internal field $\mathbf{E}^i$. The work may be performed by the field $\mathbf{E}^i$ only at the expense of the thermal energy of the semiconductor, and for this reason it is cooled. *The phenomenon in the course of which in a current-carrying conductor heat is released or absorbed due to the presence of a temperature gradient is termed Thomson effect.* Thus, *the substance is heated when the fields* $\mathbf{E}$ *and* $\mathbf{E}^i$ *are directed against each other and cooled when their directions coincide.* Thomson found that heat released in the volume dV is determined by the relation

$$dQ^T = -\tau (\nabla T \mathbf{j})\, dt\, dV; \quad dV = S\, dl, \tag{47.6}$$

where τ is the so-called *Thomson coefficient.*

There is a definite interrelation between the coefficients α_{12}, Π_{12} and τ which may be established with the aid of the laws of thermodynamics. Thermoelectric phenomena are the basic principle of operation of thermoelectric generators, thermocouples and of other devices. They play a definite part in all measurements carried out in semiconductor physics. If during the measurement the temperature of the measuring circuit contacts is not the same, the presence of the t.e.m.f. may introduce appreciable errors.

We shall present now a theoretical description of the thermoelectric phenomena based on the kinetic equation. Write a general expression for the electric current and heat flow, but in the absence of a magnetic field ($\mathbf{B}=0$):

$$\mathbf{j}=e^2K_{11}\mathbf{E}-eK_{11}T\nabla\frac{F}{T}-eK_{21}\frac{\nabla T}{T}, \qquad (47.7)$$

$$\mathbf{W}=eK_{21}\mathbf{E}-K_{21}T\nabla\frac{F}{T}-K_{31}\frac{\nabla T}{T}. \qquad (47.8)$$

Express the field $\mathbf{E}$ from (47.7) and substitute it into (47.8):

$$\mathbf{E}=\frac{1}{e^2K_{11}}\left[\mathbf{j}+eK_{11}T\nabla\frac{F}{T}+eK_{21}\frac{\nabla T}{T}\right]=$$

$$=\frac{\mathbf{j}}{e^2K_{11}}+\frac{1}{e}\nabla F+\frac{K_{21}-FK_{11}}{eK_{11}T}\nabla T. \qquad (47.9)$$

The field $\mathbf{E}$ *consists of three parts: the first term is due to the current density* $\mathbf{j}$, *i.e. it is the ohmic voltage drop, the second term is due to the inhomogeneity of the substance, and the third — to non-isothermal conditions.* Find $\mathbf{W}$ substituting (47.9) into (47.8):

$$\mathbf{W}=\frac{eK_{21}}{e^2K_{11}}\left\{\mathbf{j}+eK_{11}T\nabla\frac{F}{T}+eK_{21}\frac{\nabla T}{T}\right\}-K_{21}T\nabla\frac{F}{T}-K_{31}\frac{\nabla T}{T}=$$

$$=\frac{K_{21}}{eK_{11}}\mathbf{j}-\frac{K_{31}K_{11}-K_{21}^{2}}{TK_{11}}\nabla T. \qquad (47.10)$$

The expression (47.10) shows that energy is transported in the course of directional motion of the charge carriers resulting in the current $\mathbf{j}$ and in the course of random charge carrier motion responsible for the heat conductivity $\varkappa_e$:

$$\mathbf{W}=\frac{K_{21}}{eK_{11}}\mathbf{j}-\varkappa_e\nabla T=\Pi\mathbf{j}-\varkappa_e\nabla T. \qquad (47.11)$$

Consider again the expression for the field $\mathbf{E}$. In the absence of current in a substance ($\mathbf{j}=0$)

$$\mathbf{E}=\frac{1}{e}\nabla F+\frac{K_{21}-FK_{11}}{eTK_{11}}\nabla T. \qquad (47.12)$$

The electric field established by a temperature gradient is termed thermoelectric:

$$\mathbf{E}^{\alpha}=\frac{K_{21}-FK_{11}}{eTK_{11}}\nabla T=\alpha\nabla T. \qquad (47.13)$$

The quantity α *is termed absolute differential thermoelectromotive force:*

$$\alpha=\frac{K_{21}-FK_{11}}{eTK_{11}}. \qquad (47.13')$$

Introduce the absolute integral t.e.m.f. V_T by the condition

$$V_T = \int_0^T \alpha(\xi)\,d\xi. \tag{47.14}$$

V_T depends on the co-ordinate because so does the temperature. Find the gradient of V_T:

$$\nabla V_T = \frac{dV_I}{dT}\nabla T = \alpha(T)\nabla T = \mathbf{E}^{\alpha}. \tag{47.15}$$

Therefore, V_T is the potential of the given point with the opposite sign:

$$V_T(T(\mathbf{r})) = -\varphi(T(\mathbf{r})). \tag{47.16}$$

Defining the potential difference between the points *1* and *2* with the aid of the usual relation

$$\varphi_{12} = \varphi(1) - \varphi(2), \tag{47.17}$$

we may write

$$\varphi_{12} = V_T(2) - V_T(1). \tag{47 18}$$

Since the electric field $\mathbf{E}^{\alpha}$ is determined by a gradient of some function, i.e. it is a potential field, the circulation of $\mathbf{E}^{\alpha}$ over a closed circuit should be zero:

$$\mathscr{E} = -\oint(\mathbf{E}^{\alpha}d\mathbf{l}) = 0. \tag{47.19}$$

Hence, the e.m.f. $\mathscr{E}$ is zero. It follows from here that *it is impossible to determine the absolute thermoelectromotive force α or V_T*. The field $\mathbf{E}^{\alpha}$ determined by (47.12) *cannot establish an e.m.f. or a current in a closed circuit made of a single substance*. However, *if the circuit consists of parts made of two or more different substances, the circulation of* $\mathbf{E}^{\alpha}$ *may be non-zero and, hence, a current may flow in a closed circuit*. Consider from this aspect a circuit made up of two different substances *1* and *2* (Fig. 72). Calculate the circulation of the vector $\mathbf{E}^{\alpha}$:

$$\mathscr{E} = \mathscr{E}_{12} = -\oint(\mathbf{E}^{\alpha}\,d\mathbf{l}) = -\int_A^B(\mathbf{E}^{\alpha}\,d\mathbf{l})_1 - \int_B^A(\mathbf{E}^{\alpha}\,d\mathbf{l})_2 =$$
$$= -\int_A^B \alpha_1(\nabla T\,d\mathbf{l}) + \int_A^B \alpha_2(\nabla T\,d\mathbf{l}) = \int_{T(A)}^{T(B)}(\alpha_2 - \alpha_1)\,dT =$$
$$= \int_{T(A)}^{T(B)} \alpha_{12}\,dT. \tag{47.20}$$

The expression (47.20) coincides with (47.2). Consequently, *the relative differential thermoelectromotive force α_{12} constitutes the difference of the absolute differential thermoelectromotive forces:*

$$\alpha_{12}=\alpha_2-\alpha_1. \tag{47.21}$$

In degenerate semiconductors and in metals $K_{21}=FK_{11}$, therefore $\alpha=0$. To obtain non-zero values one should take higher terms of the expansion (3.42). In this case one obtains

$$\alpha=\frac{\pi^2}{3}\frac{k^2T}{e}\frac{\partial[\ln\sigma(E)]}{\partial E}\bigg|_{E=F}, \tag{47.22}$$

where

$$\frac{\partial[\ln\sigma(E)]}{\partial E}\bigg|_{E=F}\cong F^{-1}, \tag{47.23}$$

so that

$$\alpha\cong\frac{\pi^2}{3}\frac{k^2T}{eF}. \tag{47.24}$$

For $T\cong 300\,^\circ K$ and $F\cong 10$ eV $\alpha\cong 10^{-6}$ V/K. The relative t.e.m.f. should be of the same order of magnitude. The expression (47.24) follows from the temperature dependence of the Fermi energy. Indeed, if one takes into account that $\frac{dF}{dT}\neq 0$ and defines α by the relation

$$\alpha=\frac{1}{e}\left[\frac{dF}{dT}+\frac{K_{21}-FK_{11}}{TK_{11}}\right], \tag{47.25}$$

one obtains for $K_{21}=FK_{11}$

$$\alpha=\frac{1}{e}\frac{dF}{dT}=-\frac{\pi^2}{e}\frac{k^2T}{F}. \tag{47.26}$$

Consider now a non-degenerate semiconductor. Since, according to (46.28), $\frac{K_{21}}{K_{11}}=\frac{\langle\langle\tau\rangle\rangle}{\langle\tau\rangle}$, it follows that

$$\alpha=\frac{1}{eT}\left[\frac{\langle\langle\tau\rangle\rangle}{\langle\tau\rangle}-(F-E_c)\right]. \tag{47.27}$$

We substituted $F-E_c$ for F in (47.27) because the energy is measured from the bottom of the conduction band (or the top of the valence band in case of holes). If only one scattering mechanism is active, then, according to (46.31),

$$\frac{K_{21}}{K_{11}}=\left(\frac{5}{2}+p\right)kT$$

and

$$\alpha=\frac{k}{e}\left[\frac{5}{2}+p-\frac{F-E_c}{kT}\right]=\frac{k}{e}\left[\frac{5}{2}+p+\ln\frac{N_c}{n}\right]. \tag{47 28}$$

We see from here that α *depends in a complex way on the composition and the temperature of the semiconductor*. Since $n < N_c$, the expression in brackets is positive, and *the sign of α is determined by the sign of the charge carriers*.

Assess the value of α using the relation (47.28). Assuming that F is midway between E_c and E_d and that $\Delta E_d \cong 0.05$ eV and $2kT \cong 0.01$ eV (this corresponds to $T \cong 60°$ K) we obtain $\alpha \cong \frac{k}{e}\left[\frac{5}{2}+p+5\right]$; $\alpha \cong 9\frac{k}{e}$ for $p=3/2$. Since $\frac{k}{e}=8.62\times 10^{-5}$ V/K$=$ $=86.2\,\mu$V/K, we obtain $\alpha \cong 0.7$ mV/K, which is almost three orders of magnitude higher than the t.e.m.f. in metals.

The expression (47.28) is not valid in the very low temperature range since $\alpha \to \infty$ for $T \to 0$, while according to thermodynamics it should tend to zero, as well.

Now take the case of higher temperatures corresponding to the impurity depletion range, where $N_c > N_d$ and $n \cong N_d$. In this temperature range

$$\alpha = \frac{k}{e}\left(\frac{5}{2}+p+\ln\frac{N_c}{N_d}\right), \tag{47.29}$$

i.e. α increases with temperature.

When a semiconductor is in contact with a metal the relative t.e.m.f. will actually be equal to the t.e.m.f. of the semiconductor.

If α is defined by the relation (47.25), the expressions for t.e.m.f. in the above cases will be somewhat different.

The fact that the signs of α and of the charge carriers coincide forms the basis of the simplest method of assessing the type of conductivity, or the charge carriers' sign, with a heated probe. The material under investigation is placed on a metal plate, and a heated probe connected with the plate through a galvanometer is pressed against the material. The temperature of the material at the point of contact with the probe rises, so does the concentration of carriers, and they diffuse into the sample. An electric field is established, and the probe becomes charged with the sign corresponding to the sign of minority carriers; in n-type material the charge is positive in relation to the plate, in p-type material it is negative. A current is established in the circuit whose direction is shown by the galvanometer. Having previously established the connection between the type of charge carriers and the direction of the pointer deflection one can easily determine the sign of the charge carriers.

When evaluating α in the transition region from the impurity depletion to the intrinsic range one should remember that in the intrinsic range we will have to deal with charge carriers of two types, and that for this reason the expression derived for the case of the impurity semiconductor should be generalized for the case of intrinsic conductivity. This will be done below.

To describe the Peltier and the Thomson effects the balance of energy in the semiconductor should be considered. First of all note that there is no difference between the Peltier and the Thomson effects from the point of view of heat liberation and absorption. In both cases *energy is released as a result of work performed by internal fields in transporting the charges. In the Thomson effect this energy is generated by the thermoelectric field* $\mathbf{E}^{\alpha}=\alpha\nabla T$, *and in the Peltier effect—by the field* $\mathbf{E}^{\Pi}=-\frac{1}{e}\nabla F$ *created by the substance inhomogeneity.*

Suppose there is a temperature gradient and ∇F in a substance and, as a consequence, a thermoelectric field $\mathbf{E}^{\alpha}$. Suppose also a current of density $\mathbf{j}$ is flowing through it. The work $\dot{A}$ per unit volume per unit time performed by external and internal forces in transporting electric charge will be accompanied by heat release $\dot{Q}$ (per unit volume per unit time). Besides, the divergence of the heat flow $\mathbf{W}$ at every point should be taken into account; therefore,

$$\dot{Q}=-\operatorname{div}\mathbf{W}+\dot{A}. \tag{47.30}$$

To find $\dot{A}$ one should take note of the following. The current is induced by an external source which establishes the so-called *extraneous field* $\mathbf{E}^{ex}$. At each point the combined field $\mathbf{E}$ is equal to the sum of the internal fields $\mathbf{E}^{i}=-\frac{1}{e}\nabla F+\alpha\nabla T$ and of the extraneous field $\mathbf{E}^{ex}$.

$$\mathbf{E}=\mathbf{E}^{ex}+\mathbf{E}^{i}. \tag{47.31}$$

The current density $\mathbf{j}$ at each point should be

$$\mathbf{j}=\sigma(\mathbf{E}^{ex}+\mathbf{E}^{i})=e^2K_{11}(\mathbf{E}^{ex}+\mathbf{E}^{i}). \tag{47.32}$$

Power $\dot{A}$ released in a unit of volume is

$$\dot{A}=(\mathbf{E}\mathbf{j})=\sigma(\mathbf{E}^{ex}+\mathbf{E}^{i})^2=\sigma(\mathbf{E}^{ex})^2+\sigma\mathbf{E}^{i^2}+2\sigma(\mathbf{E}^{i}\mathbf{E}^{ex}). \tag{47.33}$$

The energy spent by the external source is, on the other hand,

$$(\mathbf{j}\mathbf{E}^{ex})=(\mathbf{E}^{ex},\ \sigma[\mathbf{E}^{ex}+\mathbf{E}^{i}])=\sigma(\mathbf{E}^{ex})^2+\sigma(\mathbf{E}^{i}\mathbf{E}^{ex}). \tag{47.34}$$

The quantity $\sigma\mathbf{E}^{i^2}+\sigma(\mathbf{E}^{i}\ \mathbf{E}^{ex})$ represents the work performed per unit time by the internal forces in transporting the charges to support the current. This work may be performed only at the expense of thermal energy If this work is positive the sample should be cooled, if it is negative the sample should be heated.

The kinetic equation should be able to describe all these processes. Let us find $\dot{Q}$. Since

$$\dot{A} = (\mathbf{E}\mathbf{j}), \tag{47.35}$$

taking into account the expiession (47.7) for j and (47.9) for $\mathbf{E}$, we may write

$$\dot{A} = (\mathbf{j}\mathbf{E}) = \left(\mathbf{j},\ \frac{\mathbf{j}}{e^2K_{11}} + \frac{1}{e}\nabla F + \frac{K_{21} - FK_{11}}{eK_{11}T}\nabla T\right) = \\ = \frac{\mathbf{j}^2}{\sigma} + \frac{1}{e}(\mathbf{j}\nabla F) + \alpha(\mathbf{j}\nabla T). \tag{47.36}$$

We see from (47.36) that the origin of the heat released is the Joule heat $\frac{\mathbf{j}^2}{\sigma}$, and the work performed by forces due to the Fermi level gradient and the temperature gradients. Taking into account that

$$\nabla F = \frac{K_{21} - FK_{11}}{TK_{11}}\nabla T = T\nabla\frac{F}{T} + \frac{K_{21}}{K_{11}}\frac{\nabla T}{T} = T\left(\nabla\frac{F}{T} - \frac{K_{21}}{K_{11}}\nabla\frac{1}{T}\right) = \\ = -T\nabla\left(\frac{K_{21} - FK_{11}}{TK_{11}}\right) = -Te\nabla\alpha, \tag{47.37}$$

we obtain for $\dot{A}$

$$\dot{A} = \frac{\mathbf{j}^2}{\sigma} - (\mathbf{j},\ T\nabla\alpha) \tag{47.38}$$

and

$$\dot{Q} = \frac{\mathbf{j}^2}{\sigma} + \operatorname{div}\varkappa_e\nabla T - (\mathbf{j},\ T\nabla\alpha). \tag{47.39}$$

Denote the Thomson heat $\dot{Q}^T$ by the relation

$$\dot{Q}^T = -\tau(\mathbf{j}\nabla T). \tag{47.40}$$

Comparing (47.40) with (47.39) we may write

$$\tau\nabla T = T\nabla\alpha. \tag{47.41}$$

Since

$$\nabla\alpha = \frac{\partial\alpha}{\partial T}\nabla T, \tag{47.42}$$

it follows that

$$\tau = T\frac{\partial\alpha}{\partial T}, \tag{47.43}$$

or

$$\alpha(T) = \int_0^T \frac{\tau(\xi)}{\xi}\,d\xi. \tag{47.44}$$

Thus, the expression (47.44) obtained on the basis of the kinetic equation connects α with τ.

Consider again the Peltier coefficient. We defined it making use of the equation

$$\mathbf{W}=\Pi\mathbf{j}-\varkappa_e\nabla T. \tag{47.45}$$

In the phenomenological theory of transport phenomena the Peltier coefficient π is usually introduced by the relation

$$\mathbf{W}=\left(\pi+\frac{\tilde{\mu}}{e}\right)\mathbf{j}-\varkappa_e\nabla T, \tag{47.46}$$

where μ is the electrochemical potential,

$$\mu=F+e\varphi, \tag{47.47}$$

and

$$\mathbf{j}=-\frac{1}{e}\sigma\nabla\tilde{\mu}-\sigma\alpha\nabla T. \tag{47.48}$$

Comparing (47.46) and (47.45) we may write

$$\Pi=\pi+\frac{F}{e}+\varphi. \tag{47.49}$$

We see from here that π *determines the kinetic energy flux, while* Π *determines the flux of the full energy, including the electrochemical potential of each charge.*

Set $\varphi=0$. In this case we have

$$\pi=\Pi-\frac{F}{e}=\frac{K_{21}}{eK_{11}}-\frac{F}{e}=\frac{1}{e}\left(\frac{K_{21}-FK_{11}}{K_{11}}\right). \tag{47.50}$$

Taking into account (47.13) we may write

$$\pi=\alpha T. \tag{47.51}$$

Consider the Peltier heat. Since (under stationary conditions) there is only the divergence of the heat flow to compensate for the work performed by the external and internal forces, and since

$$\dot{Q}=-\operatorname{div}\mathbf{W}+\dot{A}, \tag{47.52}$$

we will consider $\operatorname{div}\mathbf{W}$:

$$\operatorname{div}\mathbf{W}=\operatorname{div}\Pi\mathbf{j}-\operatorname{div}\varkappa_e\nabla T. \tag{47.53}$$

Since we are interested in the heat due to the Peltier flow we write

$$\operatorname{div}\Pi\mathbf{j}=(\nabla\Pi,\ \mathbf{j})+\Pi\operatorname{div}\mathbf{j}=(\mathbf{j},\ \nabla\Pi);\quad \operatorname{div}\mathbf{j}=0. \tag{47.54}$$

Find $\nabla\Pi$:

$$\nabla\Pi = \nabla\left(\pi + \frac{F}{e}\right) = \nabla\left(\alpha T + \frac{F}{e}\right) = T\nabla\alpha + \alpha\nabla T + \frac{1}{e}\nabla F =$$
$$= \tau\nabla T + \alpha\nabla T + \frac{1}{e}\nabla F. \qquad (47.55)$$

We see from here that *the divergence of the Peltier flow is actually related to the work performed by the fields established by the Fermi level gradient.* Usually the volume Peltier effect is of little importance, therefore we will consider only the contact effect. When current flows across the contact of two substances *the flow divergence is equal to the difference of flows*:

$$\dot{Q}^{\Pi}_{12} = \mathbf{W}_2 - \mathbf{W}_1 = \Pi_{12}\mathbf{j}_{12}. \qquad (47.56)$$

But

$$\mathbf{W}_1 = \Pi_1\mathbf{j}; \quad \mathbf{W}_2 = \Pi_2\mathbf{j}$$

and

$$\Pi_{12} = \Pi_2 - \Pi_1. \qquad (47.57)$$

The relative Peltier coefficient, i.e. the coefficient which determines the heat release at the contact of two substances when current flows from the first to the second, is equal to the difference of absolute Peltier coefficients. Thus,

$$\Pi_{12} = \Pi_2 - \Pi_1 = \left(\pi_2 + \frac{F_2}{e}\right) - \left(\pi_1 + \frac{F_1}{e}\right) =$$
$$= \pi_2 - \pi_1 + \frac{1}{e}(F_2 - F_1). \qquad (47.58)$$

For an ideal contact $F_2 = F_1$, and

$$\Pi_{12} = \pi_2 - \pi_1 = (\alpha_2 - \alpha_1)\,T = \alpha_{12}T \qquad (47.59)$$

i.e.

$$\Pi_{12} = \alpha_{12}T = \pi_{12}. \qquad (47.60)$$

In the same way other Thomson-relations for the relative thermoelectric coefficients may be determined as well.

In conclusion of the section let us consider thermoelectromotive force in a semiconductor of mixed conductivity.

Since

$$\mathbf{j} = \sum e_i K_{11(i)}(e_i\mathbf{E} - \nabla F_i) + \sum_i e_i\,(K_{11(i)}F_i - K_{21(i)})\,\frac{\nabla T}{T}, \qquad (47.61)$$

we may determine the thermoelectric field $\mathbf{E}^{\alpha}$ from the conditions $\mathbf{j}=0$ and $\nabla F_{i}=0$:

$$\mathbf{E}^{\alpha}=\frac{\sum_{i} e_{i}\left(K_{21(i)}-F_{i}K_{11(i)}\right)}{T\sum_{i} e_{i}^{2}K_{11(i)}}\nabla T=\alpha\nabla T. \qquad (47.62)$$

Let the semiconductor have both electrons and holes. Then

$$\alpha=\frac{(K_{21p}-F_{p}K_{11p})-(K_{21n}-F_{n}K_{11n})}{e_{p}T\left(K_{11p}+K_{11n}\right)}. \qquad (47.63)$$

Taking into account the expression (47.13) for t.e.m.f. for the case of charge carriers of one type we may write for (47.63)

$$\alpha=\frac{\alpha_{p}K_{11p}+\alpha_{n}K_{11n}}{K_{11p}+K_{11n}}=\frac{\alpha_{p}\sigma_{p}+\alpha_{n}\sigma_{n}}{\sigma_{p}+\sigma_{n}}. \qquad (47.64)$$

In the same way we may obtain

$$\pi=\frac{\pi_{p}\sigma_{p}+\pi_{n}\sigma_{n}}{\sigma_{p}+\sigma_{n}}. \qquad (47.65)$$

For an intrinsic semiconductor

$$\alpha=\frac{\alpha_{p}+\alpha_{n}b}{1+b}; \quad \pi=\frac{\pi_{p}+\pi_{n}b}{1+b}. \qquad (47.66)$$

The decrease of t.e.m.f. and of energy flow in a semiconductor with an intrinsic or a mixed conductivity, as compared with an extrinsic semiconductor, is due to the fact that the directions of the thermoelectric fields of the electrons and holes and their energy flows are opposite and, consequently, they attenuate each other. As this field is weakened the diffusion currents of electrons and holes grow, and this leads to a sharp increase in heat transport due to carrier recombination at the cold end, as has been already mentioned.

The choice of materials for thermocouples, thermoelectric generators, and refrigerators is based on the *efficiency parameter Z* defined by the relation

$$Z=\frac{\alpha^{2}\sigma}{\varkappa}=\frac{\alpha^{2}}{\rho\varkappa}. \qquad (47.67)$$

To understand the meaning of the quantity Z the following reasoning may be helpful. For the temperature difference ΔT the thermoelectromotive force $\mathcal{E}$ will be the greater the greater is α. To obtain a greater power from the thermoelectric generator the greatest possible current should be made to flow through it. This current is proportional to the conductivity σ and the t.e.m.f. α, and the power released will accordingly be $\alpha\sigma\alpha=\sigma\alpha^{2}$. But as the necessary temperature difference is established, a heat flow sets in the value of which for the given temperature difference ΔT is proportional to the total heat conductivity $\varkappa$. The less is $\tilde{\varkappa}$ the greater may the efficiency of the thermoelectric generator be.

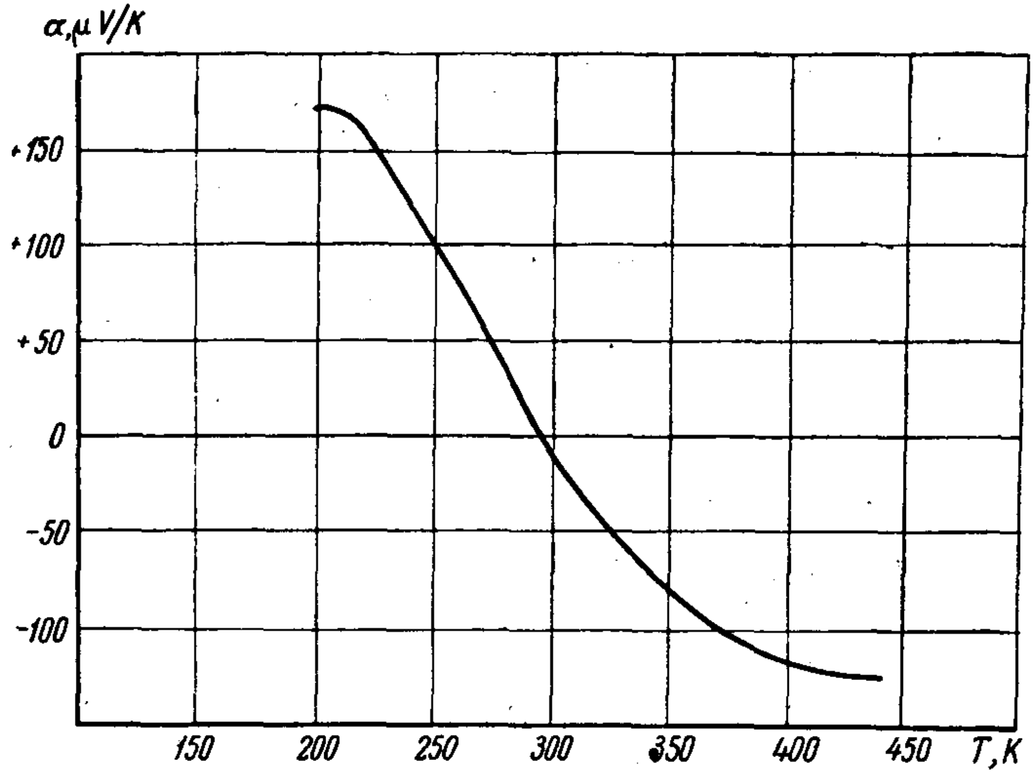


Fig 74. The temperature dependence of differential thermo-electromotive force in the copper-quicksilver telluride contact

A more precise determination of the efficiency parameter involves the use of the relative t.e.m.f. To this end, an expression is introduced for Z

$$Z=\frac{\alpha_{12}^{2}}{\left[\left(\frac{\varkappa_1}{\sigma_1}\right)^{1/2}+\left(\frac{\varkappa_2}{\sigma_2}\right)^{1/2}\right]^{2}}=\left[\frac{\alpha_2-\alpha_1}{\left(\frac{\varkappa_1}{\sigma_1}\right)^{1/2}+\left(\frac{\varkappa_2}{\sigma_2}\right)^{1/2}}\right]^{2} \qquad (47.68)$$

Figure 74 shows the temperature dependence of a sample of quicksilver telluride doped with copper. Copper, an acceptor in $A^{II}B^{VI}$-type compounds, imparts hole-type conductivity to quicksilver telluride in the low temperature range and, accordingly, a positive thermo-e.m.f. As the temperature rises the more mobile electrons begin to play the dominant part, and this leads to a change of sign of α.

Summary of Sec. 47

1. Thermoelectric phenomena involve the Seebeck, Peltier and Thomson effects.

2. In the Seebeck, or thermoelectric, effect an electric field $\mathbf{E}^{\alpha}$ is established in a sample in which a temperature gradient ∇T

has been set up:

$$\mathbf{E}^{\alpha}=\alpha\,\nabla T. \tag{47.1s}$$

The volume thermoelectric field $\mathbf{E}^{\alpha}$ prevents the separation of charges which would take place on account of the difference in charge concentrations at points of different temperature. The thermoelectric field $\mathbf{E}^{\alpha}$, by causing the conductivity current $\mathbf{j}_E=\sigma\mathbf{E}^{\alpha}$ to flow, compensates for the diffusion current $\mathbf{j}_D=-eD\,\nabla n$ caused by the concentration gradient ∇n:

$$\mathbf{j}=\mathbf{j}_E+\mathbf{j}_D=0 \quad \text{and} \quad \alpha=\frac{eD}{\sigma}\frac{dn}{dT}, \tag{47.2s}$$

where D is the diffusion coefficient.

The contact thermoelectromotive force is conditioned by the dependence of the internal contact potential difference on temperature.

3. In the extrinsic range the absolute differential thermo-e.m.f. in a non-degeneratic semiconductor

$$\alpha=\frac{k}{e}\left(\frac{5}{2}+p+\ln\frac{N_c}{n}\right). \tag{47.3s}$$

The sign of α coincides with the sign of the charge carriers. In a degenerate semiconductor

$$\alpha=\frac{\pi^2k^2}{e}\frac{T}{F}. \tag{47.4s}$$

In a semiconductor with mixed conductivity $(\sigma=\sigma_n+\sigma_p)$

$$\alpha=\frac{\alpha_n\sigma_n+\alpha_p\sigma_p}{\sigma_n+\sigma_p}. \tag{47.5s}$$

4. The absolute thermoelectromotive force α cannot be determined from experiment since the field $\mathbf{E}^{\alpha}$ cannot induce a current in a closed circuit made of a single substance because, owing to its potential nature, its circulation over this circuit is zero.

5. Thermoelectric current can flow in a circuit made up at least of two different substances. The thermoelectromotive force $\mathscr{E}_{12}$ is characterized by the relative diffe-ential t.e.m.f.

$$\alpha_{12}=\alpha_2-\alpha_1 \tag{47.6s}$$

and the temperature difference

$$\mathscr{E}_{12}=\int_{T_1}^{T_2}\alpha_{12}\,dT. \tag{47.7s}$$

6. In an inhomogeneous substance the difference in charge carrier concentrations and the separation of charges due to it leads

to the establishment of an electric field $\mathbf{E}^i$.

$$\mathbf{E}^i = -\frac{1}{e}\nabla F. \tag{47.8s}$$

When a current is flowing, the electric field $\mathbf{E}^i$ accomplishes work on transporting the charge carriers. If the directions of $\mathbf{E}^i$ and of the extraneous field $\mathbf{E}^{ex}$ responsible for the current j coincide, $\mathbf{E}^i$ will promote the current, and this will decrease the work the current source will have to perform to transport the charge. The work is performed by the field $\mathbf{E}^i$ at the expense of the thermal energy of the body causing the latter to cool. If $\mathbf{E}^i$ is directed against $\mathbf{E}^{ex}$ this will lead to an additional release of Joule heat and, consequently, the body will be heated. The amount of heat released depends on the charge that has passed through the volume (or the contact). The sign of the heat changes as the direction of the current is reversed.

Heat release resulting from the work performed by the field $\mathbf{E}^i$ (47.8s) is termed the Peltier effect. The energy released at the contact per unit area per unit time is

$$\dot{Q}_{12} = \dot{Q}^{\Pi}_{12} = \Pi_{12} j = (\Pi_2 - \Pi_1)\, j, \tag{47.9s}$$

which is equal to the difference of the Peltier energy fluxes. There is a relationship between Π and α:

$$\Pi_{12} = \alpha_{12} T = \pi_{12}, \tag{47.10s}$$

$$\pi = \Pi - \frac{F}{e} = \alpha T. \tag{47.11s}$$

7. The release or the absorption of heat $\dot{Q}^T$ in addition to the Joule heat in a substance through which a current is flowing and in which a temperature gradient ∇T exists is termed Thomson effect

$$\dot{Q}^T = -\tau(\mathbf{j}\,\nabla T), \tag{47.12s}$$

where $\dot{Q}^{Tf}$ is the amount of heat released in unit volume per unit time, and τ—Thomson coefficient. The Thomson effect is caused by the work performed by the thermoelectric field $\mathbf{E}^\alpha$ in transporting the charge in the presence of a current. There is a relationship between α and τ:

$$\tau = T\frac{\partial\alpha}{\partial T}\ ;\ \alpha(T) = \int_0^T \frac{\tau(\xi)}{\xi}\, d\xi. \tag{47.13s}$$

It follows from (47.10s) and (47.13s) that

$$\frac{d\pi_{12}}{dT} = \alpha_{12} + (\tau_2 - \tau_1) = \frac{\pi_{12}}{T} + (\tau_2 - \tau_1). \tag{47.14s}$$

8. The efficiency of a thermoelectric convertor is determined by the efficiency parameter Z:

$$Z=\left[\frac{\alpha_2-\alpha_1}{\left(\frac{\varkappa_1}{\sigma_1}\right)^{1/2}+\left(\frac{\varkappa_2}{\sigma_2}\right)^{1/2}}\right]^2. \tag{47.15s}$$

48. THERMOMAGNETIC PHENOMENA

The phenomena observed in a semiconductor with a temperature gradient set up in it in the absence of electric current in a magnetic field are termed thermomagnetic. Physically, their origin lies in the interaction of the charge carriers with the magnetic field, or, in plain words, in the linear dependence of the Lorentz force on the velocity.

The thermomagnetic phenomena include:

the Righi-Leduc effect — the appearance of a transverse temperature gradient (thermal analogue of the Hall effect);

the Maggie-Righi-Leduc effect — the variation of heat conductivity in the direction of the temperature gradient, or the appearance of a longitudinal temperature gradient;

the transverse Nernst-Ettingshausen effect — the appearance of a transverse electric field;

the longitudinal Nernst-Ettingshausen effect — the variation of the t.e.m.f., or the appearance of a longitudinal electric field.

Consider briefly the mechanism of the thermomagnetic effects.

The transverse Nernst-Ettingshausen effect, or the *thermogalvanomagnetic effect*, consists in the appearance of a transverse electric field E_z^{NE} (or the transverse potential difference V_z^{NE}) when a magnetic field $\mathbf{B}=(0, B, 0)$ is applied to the substance in which a directional energy flux $\mathbf{W}$ has been set up by a temperature gradient $\nabla T=(\nabla_x T, 0, 0)$. The intensity of the field E_z^{NE} depends on ∇T, $\mathbf{B}$ and on the properties of the substance described by the *Nernst-Ettingshausen coefficient* $A_\perp^{NE}$:

$$E_z^{NE}=A_\perp^{NE}\nabla_x T B_y. \tag{48.1}$$

For a sample of length a along the z-axis $A_\perp^{NE}$ may be determined by measuring the potential difference V_z^{NE} and applying the relation

$$A_\perp^{NE}=\frac{E_z^{NE}}{\nabla_x TB}=\frac{V_z^{NE}}{aB\nabla_x T} \tag{48.2}$$

together with the condition $\mathbf{j}=0$. The current is zero because the diffusion current is compensated by the drift current caused by the motion of the charge carriers in the thermoelectric field $\alpha\nabla T$. The drift and the diffusion currents are both affected by the magnetic field. Since the drift and the diffusion currents of particles are directed against each other, in the magnetic field they will be deflected to opposite sides. The z-component of the current, j_z, will be equal to the sum of the diffusion, j_{Dz}, and the drift, j_{Ez}, components:

$$\mathbf{j}_z^{NE}=j_{Dz}+j_{Ez}. \qquad (48.3)$$

However, the condition

$$j_x=j_{Dx}+j_{Ex}=0 \qquad (48.4)$$

does not mean that j_z turns zero

The appearance of the component j_z^{NE} results in charge accumulation on the side faces of the sample, and this, in turn, leads to the appearance of the field E_z^{NE} which compensates for the current j_z^{NE}:

$$j_z^{NE}+\sigma_{Bz}E_z^{NE}=0, \qquad (48.5)$$

whence

$$E_z^{NE}=-\frac{j_z^{NE}}{\sigma_{Bz}}=-\frac{j_{Ez}+j_{Dz}}{\sigma_{Bz}}. \qquad (48.6)$$

If **B** is reversed, E_z^{NE}, like the Hall field, will change sign. The quantity $A_\perp^{NE}$ should be dependent on the carriers' sign. Indeed, for a given ∇T the direction of the carrier flow is independent of the carriers' sign; however, the currents of electrons and holes are opposed to each other. For a given semicoductor, the sign of $A_\perp^{NE}$ will depend on which of the two components, j_{Dz} or j_{Ez}, is greater. Calculation shows *the sign of E_z^{NE} to depend on the carrier scattering mechanism*; therefore, if the type of scattering changes with temperature, E_z^{NE} may change sign.

Here is a somewhat different explanation of the thermogalvanomagnetic effect. The presence of an energy flux $\mathbf{W}_x=-\varkappa_e\nabla_x T$ in the absence of current ($\mathbf{j}=0$) means that fast particles, on the average, move against the temperature gradient, while slow ("cold") particles move along $\nabla_x T$. The Lorentz force is proportional to the velocity, and therefore the force acting on "hot" particles is greater than that acting on "cold" ones. Suppose that relaxation time is independent of energy. In this case the "hot" particles will be deflected through a greater angle than the "cold", and for this reason the oppositely directed streams of "hot" and "cold" particles will not be able to compensate each other. As a result, "hot" particles will be accumulated near the side faces of the semiconductor sample, and in this way the streams of "hot" and "cold" particles will be equalized. *The field E_z^{NE} which ensures the condition $j_z=0$ is the field of the transverse Nernst-Ettingshausen effect.* The direction of this field should be dependent on the charge carriers' sign. The above reasoning remains valid for the case when relaxation time increases with the energy as well ($\tau=\tau_0 E^p$, $p>0$). If, however, $p<0$ the "cold" particles should be deflected more than the "hot", and, in consequence, the sign of the field E_z^{NE} should be opposite to the above. In other words, *the magnitude and the sign of the field E_z^{NE} depend on the scattering mechanism.*

The longitudinal Nernst-Ettingshausen, or the longitudinal thermogalvanomagnetic, effect consists in the appearance of a longitudinal electric field, or of a potential difference, along a temperature gradient $\nabla_x T$. But since there is already a thermoelectric field $E_x^\alpha=\alpha\nabla_x T$ along $\nabla_x T$, *the appearance of an additional field along $\nabla_x T$ is tantamount to the variation of the thermoelectric field E_x^α with the applied magnetic field.* The longitudinal Nernst-Ettingshausen effect is described by the coefficient $A_\parallel^{NE}$ determined from the relation

$$E_x^\alpha(B)-E_x^\alpha(0)=[\alpha(B)-\alpha(0)]\nabla_x T=A_\parallel^{NE}\alpha(0)B^2\nabla_x T \quad (48.7)$$

whence

$$A_\parallel^{NE}=\frac{\alpha(B)-\alpha(0)}{\alpha(0)B^2}=\frac{1}{B^2}\frac{\Delta\alpha}{\alpha_0}. \quad (48.8)$$

The explanation for the longitudinal Nernst-Ettingshausen effect is quite simple. Since the relative parts played by slow and fast particles in charge transport are changed by the magnetic field, and since the action of the thermoelectric field results in energy being transported without the transfer of the charge, the application of the magnetic field should bring about a change in the thermoelectric field. The magnitude of $A_\parallel^{NE}$ should depend on the scattering mechanism, but not on the direction of the magnetic field.

The reversal of the parts played by the "hot" and the "cold" charge carriers changes the energy flux along the x-axis, i.e. creates an additional temperature gradient along the x-axis. This may be regarded as a change in heat conductivity. The variation of heat conductivity $\varkappa_e$ along the x-axis field is termed Maggie-Righi-Leduc effect. It is described by the quantity

$$\lambda=\frac{1}{B^2}\frac{\varkappa_e(0)-\varkappa_e(B)}{\varkappa_e(0)}. \quad (48.9)$$

The Righi-Leduc, or the thermomagnetic, effect (in a limited sense of the word) *consists in the appearance, on application of a magnetic field, of a transverse temperature gradient $\nabla_z T$ in a semiconductor sample with a previously established temperature gradi-*

ent $\nabla_x T$. The effect is described by the Righi-Leduc coefficient A^{RL}:

$$\nabla_z T = A^{RL} \nabla_x T B_y, \tag{48.10}$$

or

$$A^{RL} = \frac{\nabla_z T}{B \nabla_x T}. \tag{48.11}$$

Consider the mechanism of the Righi-Leduc effect. To do this we need only to revert to the explanation of the transverse Nernst-Ettingshausen effect. We concluded then that the "hot" and the "cold" carriers are deflected to opposite sides. But this means that *the face to which the "hot" carriers are deflected should be heated while the opposite face should be cooled*; in other words, *a transverse temperature difference, or gradient* $\nabla_z T$, *should appear which should cause an energy flux* $\mathbf{W}_z$ *to flow*. This energy flux would tend to compensate for the difference in energy fluxes transported by the "hot" and the "cold" particles to the side faces. When the energy fluxes along the z-axis caused by the full heat conductivity $\varkappa = \varkappa_e + \varkappa_L$ and by the magnetic field are perfectly compensated, the energy flux along the z-axis turns zero: $\mathbf{W}_z = 0$. Therefore $\mathbf{W} = (W_x, 0, 0)$ while $\nabla T = (\nabla_x T, 0, \nabla_z T)$, i.e. ∇T and $\mathbf{W}$ are set at an angle φ_{RL} to each other. According to (48.11)

$$\tan \varphi_{RL} = \frac{\nabla_z T}{\nabla_x T} = A^{RL} B. \tag{48.12}$$

It follows from (48.12) that the dimensionality of A^{RL} coincides with that of mobility. *It should depend on the charge carriers' sign and on the scattering mechanism.*

The coefficients $A_\perp^{NE}$, $A_\parallel^{NE}$, λ and A^{RL} may be calculated with the aid of the kinetic equation which takes account of the distribution of particles over states.

Summary of Sec. 48

1. In a substance in which there is an energy flux $\mathbf{W}$ caused by a temperature gradient ∇T in the absence of an electric current ($\mathbf{j} = 0$) several effects termed thermomagnetic are observed following the application of a magnetic field $\mathbf{B} = (0, B, 0)$. They include:

(a) The Righi-Leduc, or simply thermomagnetic, effect—the appearance of a transverse temperature gradient:

$$\nabla_z T = A^{RL} B \nabla_x T; \tag{48.1s}$$

(b) The transverse Nernst-Ettingshausen, or the transverse thermogalvanomagnetic, effect—the appearance of a transverse electric field

$$E_z^{NE} = A_\perp^{NE} B \nabla_x T; \tag{48.2s}$$

(c) The longitudinal Nernst-Ettingshausen, or the longitudinal thermogalvanomagnetic, effect—the appearance of an electric field along $\nabla_x T$, or the variation of t.e.m.f. $\Delta\alpha = \alpha(B) - \alpha_0$:

$$A_{\parallel}^{NE} = \frac{\Delta\alpha}{B^2\alpha_0}; \tag{48.3s}$$

(d) The Righi-Leduc-Maggie effect—the appearance of a longitudinal temperature gradient, or the variation of heat conductivity

$$\lambda = \frac{1}{B^2} \cdot \frac{\varkappa_0 - \varkappa(B)}{\varkappa_0}. \tag{48.4s}$$

The first two effects are odd, the last two—even with respect to the magnetic field.

49. GENERAL ANALYSIS OF KINETIC PHENOMENA

The description of all kinetic phenomena is contained in the kinetic equation and in the expressions (38.1s) and (38.2s) for **j** and **W** which follow from it. Those expressions should be considered together since they are determined by the same quantities; therefore, by changing conditions which govern **W** we automatically change the quantities which enter the expression for **j**, and vice versa. Here are several examples to illustrate it. Avoiding calculations to describe all the thermomagnetic phenomena and the two galvanomagnetic phenomena that have been left out of consideration we will limit ourselves to the general analysis of the equations for **j** and **W** and will demonstrate how they can be used to describe any specific effect. With this aim in view transform the expressions for current density and energy flux density.

Represent the expressions for the current density **j** and the energy flux density **W** for the case of a transverse magnetic field * as a sum of two components: the "longitudinal", $\mathbf{j}_1$, $\mathbf{W}_1$, and the "transverse", $\mathbf{j}_2$, $\mathbf{W}_2$:

$$\mathbf{j} = \mathbf{j}_1 + \mathbf{j}_2; \quad \mathbf{W} = \mathbf{W}_1 + \mathbf{W}_2, \tag{49.1}$$

*It was shown in Sec. 37 that there are no longitudinal effects in substances with scalar effective masses. The longitudinal Nernst-Ettingshausen, Maggie-Righi-Leduc, Nernst and the magnetoresistance effects are all transverse effects since they are observed in conditions of $(\nabla T\mathbf{B}) = (\mathbf{jB}) = 0$. For this reason the terms "longitudinal" and "transverse" Nernst-Ettingshausen effects are in this context not very suitable. The term "longitudinal" and "transverse" refer in this case not to the direction of **B**, but to the direction of the additional electric field (or temperature gradient) in relation to the initial fields or temperature gradients, responsible for the current or the energy flux.

where

$$\mathbf{j}_1 = eK'_{11}\left(e\mathbf{E} - T\nabla\frac{F}{T}\right) - eK'_{21}\frac{\nabla T}{T}; \tag{49.2}$$

$$\mathbf{j}_2 = \left[\frac{e^2}{m^*}K'_{12}\left(e\mathbf{E} - T\nabla\frac{F}{T}\right) - \frac{e^2}{m^*}K'_{22}\frac{\nabla T}{T},\ \mathbf{B}\right]; \tag{49.3}$$

$$\mathbf{W}_1 = K'_{21}\left(e\mathbf{E} - T\nabla\frac{F}{T}\right) - K'_{31}\frac{\nabla T}{T}; \tag{49.4}$$

$$\mathbf{W}_2 = \left[\frac{e}{m^*}K'_{22}\left(e\mathbf{E} - T\nabla\frac{F}{T}\right) - \frac{e}{m^*}K'_{32}\frac{\nabla T}{T},\ \mathbf{B}\right]. \tag{49.5}$$

Consider the component $\mathbf{j}_1$:

$$\mathbf{j}_1 = eK'_{11}(e\mathbf{E} - \nabla F) + \frac{eK'_{11}F}{T}\nabla T - \frac{eK'_{21}}{T}\nabla T =$$

$$= e^2K'_{11}\left(\mathbf{E} - \nabla\frac{F}{e}\right) - e^2K'_{11}\left(\frac{K'_{21} - FK'_{11}}{eK'_{11}T}\right)\nabla T. \tag{49.6}$$

Denote for brevity

$$e^2K'_{11} = \sigma_B. \tag{49.7}$$

Taking into account that

$$\frac{K'_{21} - FK'_{11}}{eK'_{11}T} = \alpha(\mathbf{B}) = \alpha', \tag{49.8}$$

the component $\mathbf{j}_1$ may be written in the form

$$\mathbf{j}_1 = \sigma_B\left\{\left(\mathbf{E} - \nabla\frac{F}{e}\right) - \alpha'\nabla T\right\}. \tag{49.9}$$

Transform the first multiplier of $\mathbf{j}_2$ as follows:

$$\frac{e^2}{m^*}K'_{12}\left(e\mathbf{E} - T\nabla\frac{F}{T}\right) - \frac{e^2}{m^*}K'_{22}\frac{\nabla T}{T} =$$

$$= e^2K'_{11}\frac{eK'_{12}}{m^*K'_{11}}\left(\mathbf{E} - \nabla\frac{F}{e}\right) + \frac{e^2}{m^*}\frac{K'_{12}F - K'_{22}}{T}\nabla T. \tag{49.10}$$

Taking into account that $\frac{e}{m^*}\frac{K'_{12}}{K'_{11}} = \mu^H$ is the Hall mobility, and denoting by analogy with (49.8)

$$\frac{K'_{22} - FK'_{12}}{eTK'_{12}} = \beta(\mathbf{B}) = \beta', \tag{49.11}$$

we may write

$$\frac{e^2}{m^*}\frac{K'_{12}F - K'_{22}}{T} = e^2K'_{11}\frac{e}{m^*}\frac{K'_{12}}{K'_{11}}\left(\frac{K'_{12}F - K'_{22}}{eTK'_{12}}\right) = -\sigma_B\mu^H\beta', \tag{49.12}$$

and

$$\mathbf{j}_2=\sigma_B\mu^H\left[\left(\mathbf{E}-\nabla\frac{F}{e}\right)-\beta'\nabla T,\ \mathbf{B}\right]. \tag{49.13}$$

In the same way transform the expression for the energy flux. Write the first term of $\mathbf{W}_1$ as follows:

$$\begin{aligned}\mathbf{W}_1&=K'_{21}(e\mathbf{E}-\nabla F)+\frac{FK'_{21}}{T}\nabla T-K'_{31}\frac{\nabla T}{T}=\\&=e^2K'_{11}\frac{K'_{21}}{eK'_{11}}\left(\mathbf{E}-\nabla\frac{F}{e}\right)-\frac{K'_{31}-FK'_{21}}{T}\nabla T=\\&=\sigma_B\left(E-\nabla\frac{F}{e}\right)\Pi-\frac{K'_{31}-FK'_{21}}{T}\nabla T.\end{aligned} \tag{49.14}$$

But

$$\begin{aligned}K'_{31}-FK'_{21}&=K'_{31}-\frac{K'^2_{21}}{K'_{11}}+\frac{K'^2_{21}}{K'_{11}}-FK'_{21}=\\&=\frac{K'_{31}K'_{11}-K'^2_{21}}{K'_{11}}+K'_{21}\left(\frac{K'_{21}-FK'_{11}}{K'_{11}}\right)=\varkappa'_eT+K'_{21}eT\alpha'=\\&=\varkappa'_eT+\Pi e^2K'_{11}T\alpha'=\varkappa'_eT+\sigma_B\Pi\alpha'T,\end{aligned} \tag{49.15}$$

therefore,

$$\mathbf{W}_1=\sigma_B\left(\mathbf{E}-\nabla\frac{F}{e}-\alpha'\nabla T\right)\Pi-\varkappa'_e\nabla T. \tag{49.16}$$

Consider the first multiplier of $\mathbf{W}_2$:

$$\begin{aligned}&\frac{e}{m^*}K'_{22}\left(e\mathbf{E}-T\nabla\frac{F}{T}\right)-\frac{e}{m^*}K'_{32}\frac{\nabla T}{T}=\frac{e^2}{m^*}K'_{22}\left(\mathbf{E}-\nabla\frac{F}{e}\right)+\\&+\frac{e}{m^*}\left(\frac{FK'_{22}-K'_{32}}{T}\right)\nabla T=e^2K'_{11}\frac{K'_{21}}{eK'_{11}}\frac{eK'_{22}}{m^*K'_{21}}\left(\mathbf{E}-\nabla\frac{F}{e}\right)+\\&+\frac{e}{m^*}\left(\frac{FK'_{22}-K'_{32}}{T}\right)\nabla T=\sigma_B\Pi\mu^E\left(\mathbf{E}-\nabla\frac{F}{e}\right)+\\&+\frac{e}{m^*}\frac{FK'_{22}-K'_{32}}{T}\nabla T.\end{aligned} \tag{49.17}$$

μ^E denotes the quantity

$$\mu^E=\frac{e}{m^*}\frac{K'_{22}}{K'_{21}}=\frac{e}{m^*}\frac{\left\langle\!\left\langle\dfrac{\tau^2}{1+\mu^2B^2}\right\rangle\!\right\rangle}{\left\langle\!\left\langle\dfrac{\tau}{1+\mu^2B^2}\right\rangle\!\right\rangle}. \tag{49.18}$$

Transform the addend in expression (49.17):

$$\begin{aligned}K'_{32}-FK'_{22}&=K'_{32}-\frac{K'^2_{22}}{K'_{12}}+\frac{K'^2_{22}}{K'_{12}}-K'_{22}F=\\&=\frac{K'_{32}K'_{12}-K'^2_{22}}{K'_{12}}+\frac{K'_{22}(K'_{22}-FK'_{12})}{K'_{12}}.\end{aligned} \tag{49.19}$$

The second term of (49.19) may be written as follows:

$$\frac{K'_{22}(K'_{22}-FK'_{12})}{K'_{12}}=K'_{22}eT\beta'=\frac{e}{m^*}\frac{K'_{22}}{K'_{21}}\frac{K'_{21}}{eK'_{11}}em^*K'_{11}T\beta'=$$

$$=\sigma_B\mu^E\Pi T\beta'\frac{m^*}{e}. \qquad (49.20)$$

Taking into account $\frac{e}{m^*T}$ write the first addend in the last term of (49.19) in the form

$$\frac{e}{m^*T}\frac{K'_{32}K'_{12}-K'^2_{22}}{K'_{12}}=\frac{eK'_{22}}{m^*K'_{21}}\frac{K'_{21}}{K'_{22}}\frac{K'_{32}K'_{12}-K'^2_{22}}{TK'_{12}}=\mu^E\lambda'. \qquad (49.21)$$

Making use of (49.17, 18, 20 and 21) we write

$$\mathbf{W}_2=\left[\sigma_B\mu^E\left(\mathbf{E}-\nabla\frac{F}{e}-\beta'\nabla T\right)\Pi-\mu^E\lambda'\nabla T,\ \mathbf{B}\right]. \qquad (49.22)$$

Denoting

$$\mathbf{E}-\nabla\frac{F}{e}=\mathbf{E}^* \qquad (49.23)$$

re-write the expressions for the current density and the energy flux:

$$\mathbf{j}=\sigma_B(\mathbf{E}^*-\alpha'\nabla T)+\sigma_B\mu^H[\mathbf{E}^*-\beta'\nabla T,\ \mathbf{B}] \qquad (49.24)$$

and

$$\mathbf{W}=\sigma_B(\mathbf{E}^*-\alpha'\nabla T)\Pi-\varkappa'_e\nabla T+$$

$$+\mu^E[\sigma_B(\mathbf{E}^*-\beta'\nabla T)\Pi-\lambda'\nabla T,\ \mathbf{B}]. \qquad (49.25)$$

The equations (49.24) and (49.25) describe all phenomena connected with charge carrier motion, namely:

electric and heat conductivities;
thermoelectric phenomena;
galvanomagnetic phenomena;
thermomagnetic phenomena.

All these effects take place simultaneously being superimposed one upon the other. To separate one specific effect in its "pure" form certain experimental conditions should be created, otherwise several effects may be observed at a time. Below we will describe the conditions in which a certain effect may be observed individually. The equations (49.24) and (49.25) describe both isothermal and adiabatic effects. Now we will point out the meaning of the terms of the equations (49.24) and (49.25).

The term $\sigma_B(\mathbf{E}^*-\alpha'\nabla T)$ describes the conduction current in the combined electric field established by electrostatic potential gradient, the Fermi energy gradient and the t.e.m.f.; $\sigma_B\mu^H[\mathbf{E}^*\mathbf{B}]$ is the Hall current, and the term

$$-\sigma_B\mu^H\beta'[\nabla T,\mathbf{B}] \qquad (49.26)$$

describes the current which it would be natural to term Nernst-Ettingshausen current, since it is this current which is responsible for the appearance of the Nernst-Ettingshausen transverse electric field. Since the values of the conductivity and of the t.e.m.f. depend on **B**, the equations (49.24) and (49.25) contain the description of the magnetoresistance and the longitudinal Nernst-Ettingshausen effects.

The expression for the energy flux consists of terms of two types. In the terms of one type the Peltier coefficient enters as a multiplier; obviously, these terms describe phenomena connected with energy transport in the presence of a current and lead to galvanomagnetic effects. The terms containing ∇T and **B** are responsible for the thermomagnetic phenomena.

Now we will demonstrate how the equations (49.24) and (49.25) should be used to describe specific effects.

I. Isothermal effects: $\nabla T = 0$.

1. *In the absence of a magnetic field* ($\mathbf{B}=0$) the equations (49.24) and (49.25) assume the form

$$\mathbf{j}=\sigma_0\mathbf{E}^*; \quad \mathbf{W}=\sigma_0\mathbf{E}^*\Pi=\Pi\mathbf{j}. \tag{49.27}$$

They describe the conduction current and the Peltier flux caused by the electric field $\mathbf{E}^*=-\nabla\left(\varphi+\frac{F}{e}\right)$. In a homogeneous substance ($\nabla F=0$) the current is caused by the gradient of the electrostatic potential, in an inhomogeneous one, by the gradient of the electrochemical potential. If in the latter case $\mathbf{j}=0$ there will be a gradient of electrostatic potential in the sample

$$-\nabla\varphi=\nabla\frac{F}{e}; \quad \mathbf{E}=\frac{1}{e}\nabla F. \tag{49.28}$$

The equation (49.28) discloses a new phenomenon: *in an inhomogeneous semiconductor there is a volume electric field.*

2. *Semiconductor is in magnetic fields.* In this case the equations for **j** and **W** will be of the form

$$\mathbf{j}=\sigma_B\mathbf{E}^*+\sigma_B\mu^H[\mathbf{E}^*\mathbf{B}], \tag{49.29}$$

$$\mathbf{W}=\sigma_B\mathbf{E}^*\Pi+\sigma_B\mu^E[\mathbf{E}^*\mathbf{B}]\Pi. \tag{49.30}$$

The first term in **j** describes the ohmic current, the second—the Hall current, Peltier heat currents being connected with them. By setting definite boundary conditions one may arrive at the isothermal Hall effect. To achieve this end it suffices to assume

$$\mathbf{j}=(j_x,\ 0,\ 0); \quad \mathbf{B}=(0,\ B,\ 0); \quad \mathbf{E}^*=(E_x^*.\ 0,\ E_z^*). \tag{49.31}$$

The expression for the flux **W** may not be chosen arbitrarily; for instance, it may not be assumed that $\mathbf{W}=(W_x,\ 0,\ 0)$ since the expressions for **j** and **W** will become incompatible.

We should put $\mathbf{W} = (W_x, 0, W_z)$ and write for W_z

$$W_z = \sigma_B E_z^* \Pi + \sigma_B \mu^E E_x^* B \Pi. \qquad (49.32)$$

At the same time

$$j_z = \sigma_B E_z^* + \sigma_B \mu^H E_x B. \qquad (49.33)$$

Indeed, $W_z = 0$ does not follow from $j_z = 0$. This can be easily explained from physical considerations: j_z turns zero because a charge is accumulated on the side faces of the sample, which establishes a Hall field, and it, in turn, compensates the magnetic field. It is not the same with the heat flux. There are no heat sources to compensate for the heat flux W_z, therefore, there is an ideal thermal contact between the sample and the ambient which ensures the continuity of the flux W_z, and this, in turn, ensures that the condition $\nabla T = 0$ is satisfied.

If W_z is required to be zero, the heat flux in the OZ direction caused by the magnetic field will have to be compensated by the heat conductivity flux; consequently, a temperature gradient in the OZ direction will have to be established. We see from here that *the Ettingshausen effect is only possible if there is no energy exchange between the sample and the ambient*, i.e. the Ettingshausen effect is an adiabatic one. Before we discuss it consider other effects. The equations (49.29) and (49.30) contain the description of the *isothermal magnetoresistance* both for a finite (together with 49.31) and for an infinite ($E_z = 0$, $j_z \neq 0$) sample.

II. Non-isothermal effects: $\nabla T \neq 0$.

1. *In the absence of a magnetic field* ($\mathbf{B} = 0$) the equations (49.24) and (49.25) assume the form

$$\mathbf{j} = \sigma_0 (\mathbf{E}^* - \alpha \nabla T), \qquad (49.34)$$

$$\mathbf{W} = \sigma_0 (\mathbf{E}^* - \alpha \nabla T) \Pi - \varkappa_e \nabla T = \Pi \mathbf{j} - \varkappa_e \nabla T. \qquad (49.35)$$

The condition $\mathbf{j} = 0$ determines a heat conductivity flux

$$\mathbf{W} = -\varkappa_e \nabla T. \qquad (49.36)$$

A thermoelectric field

$$\mathbf{E}^* = \alpha \nabla T = \mathbf{E}^\alpha \qquad (49.37)$$

is established in the substance.

If $\mathbf{j} \neq 0$ the conditions for all the thermoelectric phenomena discussed above (the Seebeck, Peltier and Thomson effects) will be satisfied.

2. *Semiconductor is in a magnetic field:* $\mathbf{B} \neq 0$. The general expressions (49.24) and (49.25) determine the current $\mathbf{j}$ and the energy flux $\mathbf{W}$.

If prior to the application of the magnetic field the current flowed in the x-axis direction and ∇T was zero, the condition

$\nabla T=0$ may be changed by the magnetic field in two ways:

$$\nabla T=(0,\ 0,\ \nabla_z T), \tag{49.38}$$

$$\nabla T=(\nabla_x T,\ 0,\ 0). \tag{49.39}$$

The first (49.38) corresponds to the Ettingshausen effect, the second (49.39)—to the Nernst effect. Now we will illustrate the application of the equations (49.34) and (49.35) with the example of the Ettingshausen effect.

Ettingshausen effect. To observe it one should put

$$\mathbf{j}=(j_x,\ 0,\ 0);\ \mathbf{W}=(W_x,\ 0,\ 0). \tag{49.40}$$

As was already mentioned, for the conditions $j_z=W_z=0$ to be satisfied simultaneously, it is necessary that $\nabla_z T\neq 0$; therefore we will presume

$$\nabla T=(0,\ 0,\ \nabla_z T). \tag{49.41}$$

Write the expressions (49.24) and (49.25) for $\mathbf{j}$ and $\mathbf{W}$ taking into account (49.40) and (49.41); $\mathbf{B}=(0,\ B,\ 0)$:

$$j_x=\sigma_B E_x^* - \sigma_B\mu^H(E_z^* - \beta'\nabla_z T)B; \tag{49.42}$$

$$j_z=\sigma_B(E_z^* - \alpha'\nabla_z T)+\sigma_B\mu^H E_x^* B=0; \tag{49.43}$$

$$W_x=\sigma_B E_x^*\Pi-\mu^E(\sigma_B E_z^*\Pi-\sigma_B\beta'\nabla_z T\Pi-\lambda'\nabla_z T)B; \tag{49.44}$$

$$W_z=\sigma_B(E_z^*-\alpha'\nabla_z T)\Pi+\mu^E\sigma_B E_x^*\Pi B-\varkappa_e'\nabla_z T=0. \tag{49.45}$$

Find E_z^* and $\nabla_z T$ from the conditions $j_z=W_z=0$. From (49.43) we write

$$\sigma_B(E_z^*-\alpha'\nabla_z T)=-\sigma_B\mu^H E_x^* B. \tag{49.46}$$

Substitute (49.46) into (49.45) to obtain

$$W_z=-\sigma_B\mu^H E_x^*\Pi B+\mu^E\sigma_B E_x^*\Pi B-\varkappa_e'\nabla_z T=0. \tag{49.47}$$

From (49.47) we find

$$\nabla_z T=\frac{(\mu^E-\mu^H)}{\varkappa_e'}\sigma_B E_x^* B\Pi. \tag{49.48}$$

Substituting $\nabla_z T$ of (49.48) into (49.46), find E_z^*:

$$E_z^*=\alpha'\nabla_z T-\mu^H E_x^* B=\left(\frac{\alpha'\sigma_B\Pi\,(\mu^E-\mu^H)}{\varkappa_e'}-\mu^H\right)E_x^* B. \tag{49.49}$$

The expression (49.49) shows that in the adiabatic case ($W_z=0$) a thermoelectric field $\alpha'\nabla_z T$ is added to the isothermal Hall field ($-\mu^H E_x^* B$); therefore, *the adiabatic Hall coefficient R_a is not the same as the isothermal one, R_I*. To introduce the Ettingshausen A^E and the Hall R_a coefficients, $\nabla_z T$ and E_z^* should be expressed in terms of j_x and B. To this end substitute the expressions obtained by us for E_z^* and $\nabla_z T$ into (49.42):

$$j_x = \sigma_B E_x^* - \sigma_B \mu^H \left\{ \frac{\sigma_B \alpha' (\mu^E - \mu^H) \Pi}{\varkappa_e'} - \mu^H - \frac{\beta' \sigma_B (\mu^E - \mu^H) \Pi}{\varkappa_e'} \right\} B E_x^* =$$

$$= \left\{ \sigma_B - \sigma_B \mu^H B \left[\frac{\sigma_B \Pi (\mu^E - \mu^H)(\alpha' - \beta')}{\varkappa_e'} - \mu^H \right] \right\} E_x^* = \sigma_{Bx} E_x^*. \qquad (49.50)$$

Find the coefficients A^E and R_a from (49.48) and (49.49) taking into account (49.50):

$$A^E = \frac{\nabla_z T}{j_x B} = \frac{\sigma_B \Pi (\mu^E - \mu^H) E_x^* B}{\varkappa_e' B j_x} = \frac{\sigma_B \Pi (\mu^E - \mu^H)}{\varkappa_e' \sigma_{Bx}} \qquad (49.51)$$

and

$$R_a = -\frac{E_z^*}{j_x B} = \frac{\left[\mu^H - \frac{\alpha' \sigma_B \Pi (\mu^E - \mu^H)}{\varkappa_e'} \right] E_x^* B}{B j_x} =$$

$$= \left[\frac{\mu^H}{\sigma_{Bx}} - \frac{\alpha' \sigma_B \Pi (\mu^E - \mu^H)}{\varkappa_e' \sigma_{Bx}} \right] = \frac{\mu^H}{\sigma_{Bx}} - \alpha' A^E = R_I - \alpha' A^E. \qquad (49.52)$$

The quantity σ_{Bx} as determined from (49.51) describes the *adiabatic magnetoresistance*. Find A^E and R_a for weak fields (for $B \rightarrow 0$). In this case $\sigma_B \cong \sigma_0$; $\sigma_{Bx} \cong \sigma_0$, and

$$A^E = \frac{\Pi (\mu^E - \mu^H)}{\varkappa_e}, \qquad (49.53)$$

$$R_a = R_I - \alpha A^E. \qquad (49.54)$$

The meaning of the difference in R_a and R_I is obvious. *The field E_z^* is in adiabatic conditions weaker than in isothermal conditions by the thermoelectric field $\alpha \nabla_z T$*, and *$\nabla_z T$ corresponding to a unit current and a unit magnetic field induction exactly equals A^E.*

Calculate A^E. Note first of all that the flux W_z will be determined not only by the electron heat conductivity but by phonon heat conductivity as well. Therefore full heat conductivity $\varkappa = \varkappa_e + \varkappa_L$ should be substituted for $\varkappa_e$:

$$A^E = \frac{\Pi (\mu^E - \mu^H)}{\varkappa} = \frac{1}{\varkappa} \frac{K_{21}}{e K_{11}} \left(\frac{e}{m^*} \frac{K_{22}}{K_{21}} - \frac{e}{m^*} \frac{K_{12}}{K_{11}} \right) =$$

$$= \frac{e}{m^*} \frac{1}{\varkappa e} \cdot \frac{\langle\langle \tau \rangle\rangle}{\langle \tau \rangle} \left(\frac{\langle\langle \tau^2 \rangle\rangle}{\langle\langle \tau \rangle\rangle} - \frac{\langle \tau^2 \rangle}{\langle \tau \rangle} \right) = \frac{\mu_d}{e\varkappa} \frac{\langle\langle \tau \rangle\rangle}{\langle \tau \rangle} \left(\frac{\langle\langle \tau^2 \rangle\rangle}{\langle\langle \tau \rangle\rangle \langle \tau \rangle} - \frac{\langle \tau^2 \rangle}{\langle \tau \rangle^2} \right). \qquad (49.55)$$

The expression (49.55) shows that *the Ettingshausen effect is independent of the charge carriers' sign, but depends on the scattering* *mechanism*. Indeed, let $\tau = \tau_0 E^p$. In this case

$$\langle\tau\rangle = \tau_0' \frac{\Gamma\left(\frac{5}{2}+p\right)}{\Gamma\left(\frac{5}{2}\right)}; \quad \langle\tau^2\rangle = \tau_0'^2 \frac{\Gamma\left(\frac{5}{2}+2p\right)}{\Gamma\left(\frac{5}{2}\right)};$$

$$\langle\langle\tau\rangle\rangle = \frac{5}{2}(kT)\,\tau_0' \frac{\Gamma\left(\frac{7}{2}+p\right)}{\Gamma\left(\frac{7}{2}\right)}; \quad \langle\langle\tau^2\rangle\rangle = \frac{5}{2}(kT)\,\tau_0'^2 \frac{\Gamma\left(\frac{7}{2}+2p\right)}{\Gamma\left(\frac{7}{2}\right)}. \qquad (49.56)$$

Substitute the values of the averaged relaxation times (49.56) into (49.55) to obtain

$$A^E = \frac{\mu_d}{e\varkappa} \frac{\frac{5}{2}(kT)\Gamma\left(\frac{7}{2}+p\right)\Gamma\left(\frac{5}{2}\right)}{\Gamma\left(\frac{7}{2}\right)\Gamma\left(\frac{5}{2}+p\right)} \left[\frac{\frac{5}{2}(kT)\Gamma\left(\frac{7}{2}+2p\right)\Gamma\left(\frac{7}{2}\right)\Gamma\left(\frac{5}{2}\right)}{\Gamma\left(\frac{7}{2}\right)\frac{5}{2}(kT)\Gamma\left(\frac{7}{2}+p\right)\Gamma\left(\frac{5}{2}+p\right)} - \right.$$

$$\left. - \frac{\Gamma\left(\frac{5}{2}+2p\right)\left[\Gamma\left(\frac{5}{2}\right)\right]^2}{\Gamma\left(\frac{5}{2}\right)\left[\Gamma\left(\frac{5}{2}+p\right)\right]^2}\right] = p\,\frac{kT\mu_d}{\varkappa e} \frac{\Gamma\left(\frac{5}{2}+2p\right)\Gamma\left(\frac{5}{2}\right)}{\left[\Gamma\left(\frac{5}{2}+p\right)\right]^2} =$$

$$= p\,\frac{kT}{e\varkappa}\,\mu^H = p\,\frac{kT}{\varkappa}\,\frac{R_l\sigma_0}{e}. \qquad (49.57)$$

For $p > 0$ the Ettingshausen coefficient $A^E > 0$, i.e. $\nabla_z T > 0$ for $B > 0$ and $j_x > 0$—the upper face is heated and the lower cooled. This means that the upward deflection of "hot" carriers is greater. When $p = 0$ relaxation time is independent of energy, the deflection is not selective, and a temperature gradient is not set up. When $p < 0$, $A^E < 0$, and "hot" carriers are deflected downwards (for the chosen directions of fields and of the current). If, as the result of changes in the composition, the structure or the temperature of the sample, its Ettingshausen effect changes signs, this means that the respective parts played by different scattering mechanisms have changed, i.e. that one mechanism has been superceded by another.

Assess the value of A^E. Let $\varkappa = 1$ W·cm/(cm²·K) $= 10^2$ W (m·K)$^{-1}$, $kT = 0.03$ eV; $\mu^H = 10\,000$ cm²/(V·s). Then A^E will be equal to (—) 150 cm³·K/J = (—)1.5×10⁻⁴ m³·K/J and 4.5×10⁻⁴ m³·K/J for $p = $ (—)1/2 and 3/2, respectively. For a current density of $j_x = 1$ A/cm² $= 10^4$ A/m² and $B = 1$ $T = 10^4$ Gs the values of $\nabla_z T = A^E j_x B$ are (—) 1.5 K/m and 4.5 K/m $= 0.05$ K/cm. As you see, the gradient is comparatively small. Assess αA^E: for $\alpha = 1$ mV/K $= 10^{-3}$ V/K, $p = 3/2$; the value of $\alpha A^E = 5\times10^{-7}$ m³·K×V/(J·K) $= 5\times10^{-7}$ m³/C. For lightly doped semiconductors this correction may be neglected, but *for heavily doped semiconductors the difference between R_a and R_l should be taken into account.* If one

Table 15

Conditions in Which Transverse Galvanomagnetic and Thermomagnetic Effects are Observed

Galvanomagnetic effects	Formula for the coefficient	Thermomagnetic effects	Formula for the coefficient
Hall effect — transverse field E_z $(j_z = \nabla_x T = 0)$	$R = -\frac{E_z}{j_x B_y}$	Righi-Leduc effect — transverse temperature gradient $\nabla_z T$ $(\mathbf{j} = 0)$	$A^{RL} = \frac{\nabla_z T}{B_y \nabla_x T}$
Magnetoresistance — resistance variations, longitudinal potential difference ΔV_x $(j_z = \nabla_x T = 0)$	$H = \frac{1}{B^2} \frac{\rho(B) - \rho(0)}{\rho(0)}$	Maggie-Righi-Leduc effect — thermal conductivity variations, longitudinal temperature difference $\Delta(\nabla_x T)$ $(\mathbf{j} = 0)$	$\lambda = \frac{1}{B^2} \frac{\varkappa(B) - \varkappa(0)}{\varkappa(0)}$
Ettingshausen effect — transverse temperature gradient $\nabla_z T$ $(j_z = \nabla_x T = 0)$	$A^E = \frac{\nabla_z T}{j_x B_y}$	Transverse Nernst-Ettingshausen effect — transverse electric field E_z $(\mathbf{j} = 0)$	$A_{\perp}^{NE} = \frac{E_z}{B_y \nabla_x T}$
Nernst effect — longitudinal temperature difference, longitudinal temperature gradient $\nabla_x T$ $(j_z = 0)$	$A^N = \frac{\nabla_x T}{j_x B_y^2}$	Longitudinal Nernst-Ettingshausen effect — the appearance of an electric field and a potential difference, variations of thermo-e.m.f. $\Delta(E_x)$ $(\mathbf{j} = 0)$	$A_{\parallel}^{NE} = \frac{1}{B^2} \frac{\alpha(B) - \alpha(0)}{\alpha(0)}$

Isothermic effects: $\nabla_z T = 0$; adiabatic effects: $W_z = 0$. Magnetic field $\mathbf{B} = (0, B, 0)$.

assumes $\nabla_x T \neq 0$ he may obtain the expression for the Nernst effect. Two effects are possible: the isothermal if $\nabla_z T = 0$ and the adiabatic if $W_z = 0$. A^N is calculated in the same way as A^E.

III. Thermomagnetic phenomena.

To describe thermomagnetic phenomena one should put

$$\mathbf{j} = 0; \quad \nabla T \neq 0, \tag{49.58}$$

therefore,

$$j_x = \sigma_B (E_z^* - \alpha' \nabla_x T) - \sigma_B \mu^H (E_z^* - \beta' \nabla_z T) B = 0, \tag{49.59}$$

$$j_z = \sigma_B (E_z^* - \alpha' \nabla_z T) + \sigma_B \mu^H (E_x^* - \beta' \nabla_x T) B = 0 \tag{49.60}$$

$$W_x = \sigma_B (E_x^* - \alpha' \nabla_x T) \Pi - \varkappa_e' \nabla_x T - \mu^E \sigma_B (E_z^* - \beta' \nabla_z T) B\Pi + \\ + \mu^E \lambda' \nabla_z T \cdot B, \tag{49.61}$$

$$W_z = \sigma_B (E_z^* - \alpha' \nabla_z T) \Pi - \varkappa_e' \nabla_z T + \\ + \mu^E \sigma_B (E_x^* - \beta' \nabla_x T) B\Pi - \mu^E \lambda' \nabla_x T B. \tag{49.62}$$

To obtain the expressions for the coefficients of the longitudinal and the transverse Nernst-Ettingshausen and Righi-Leduc effects one should find the values of E_x^*, E_z^*, $\nabla_z T$ from the equations (49.59-62) and express them in terms of B and $\nabla_x T$. Evidently, the expressions for E_x^* and E_z^* may be sought for two different assumptions: $\nabla_z T = 0$ *for the isothermal effects, or* $W_z = 0$ *for the adiabatic effects. The Righi-Leduc effect is always adiabatic.* Having found E_x^*, E_z^* and $\nabla_z T$ and substituted them into the expression for W_x we obtain $W_x = -\varkappa_e' \nabla_x T$ whence follows the expression for the Maggie-Righi-Leduc effect ($(\varkappa_e + \varkappa_L) \nabla_z T$ should be substituted for $\varkappa_e' \nabla_z T$).

Taking into account the magnetic field dependence of kinetic coefficients one may find the coefficients of galvano- and thermomagnetic effects both for weak and for strong fields in both the extrinsic and intrinsic conductivity ranges.

Summary of Sec. 49

1. The expressions (49.24) and (49.25) for **j** and **W** contain the description of all the kinetic phenomena in semiconductors with scalar effective masses. Table 15 presents the conditions in which specific effects should be observed.

50. ON KINETIC PHENOMENA IN SEMICONDUCTORS WITH TENSOR EFFECTIVE MASSES

In Sec. 40 we discussed the electric conductivity in semiconductors with tensor effective masses. The application of a magnetic field appreciably complicates the expression for the current. Consider first the quantity (**B**, **m*** **B**). Let the components of **B** in the main axes of the tensor $\mathbf{m}^*$ (or $\mathbf{m}^{*-1}$, or $\sigma^{(\nu)}$) be (B_1, B_2, B_3). In this case the components of the vector $\mathbf{m}^*\mathbf{B}$ will be

$$\mathbf{m}^*\mathbf{B} = (m_1B_1,\ m_2B_2,\ m_3B_3) \tag{50.1}$$

and the scalar product $(\mathbf{B}, \mathbf{m}^*\mathbf{B})$ will, therefore, depend on the direction of $\mathbf{B}$ relative to the main axes of the ellipsoid $\mathbf{m}^*$.

$$(\mathbf{B}, \mathbf{m}^*\mathbf{B}) = m_1B_1^2 + m_2B_2^2 + m_3B_3^2 = m_t(B_1^2 + B_2^2) + m_lB_3^2. \tag{50.2}$$

It may change from m_lB^2 when $\mathbf{B}$ is directed along the energy ellipsoid rotational axis to m_tB^2 when $\mathbf{B}$ lies in a plane perpendicular to this axis. But this means that the kinetic coefficient K'_{rs} depends on the magnetic field $\mathbf{B}$.

Defining the direction of $\mathbf{B}$ relative to the rotational axis by the angle θ we see that

$$(\mathbf{B}, \mathbf{m}^*\mathbf{B}) = B^2(m_t \sin^2\theta + m_l \cos^2\theta) \tag{50.3}$$

is a periodic function of θ.

Write the expression (38.7) for $\mathbf{j}^{(\nu)}$:

$$\mathbf{j}^{(\nu)} = e^2K_{11}'^{(\nu)}\mathbf{E} + e^3K_{12}'^{(\nu)}[\mathbf{m}^{*-1}\mathbf{E},\ \mathbf{B}] + \frac{e^4}{|\mathbf{m}^*|}K_{13}'^{(\nu)}(\mathbf{E}\mathbf{B})\,\mathbf{m}^*\mathbf{B}. \tag{50.4}$$

We see that the expression (50.4) for the current is much more complex as compared with the case of $\mathbf{B} = 0$. We shall not analyze this expression, or discuss kinetic phenomena on the basis of the equation (50.4). In principle, it presents no difficulties. Just write the equation (50.4) in tensor form. As we have already seen, the first term represents a tensor of rank II $\sigma_{ij}^{(\nu)}$. The second term may be represented in the form of a rank III tensor $\sigma_{ijk}'^{(\nu)}$, and the third enables us to introduce a tensor of rank IV $\sigma_{ijkl}'^{(\nu)}$, the components of all the tensors being dependent on the magnetic field. For a weak field in which $\frac{e^2\tau^2}{|\mathbf{m}^*|}(\mathbf{B}, \mathbf{m}^*\mathbf{B}) \ll 1$ the above tensors will be independent of the magnetic field, and $\mathbf{j}^{(\nu)}$ may be expressed in the form

$$j_i^{(\nu)} = \sum_{l=1}^{3}\sigma_{il}^{(\nu)}E_l + \sum_{l,\,k=1}^{3}\sigma_{ilk}^{(\nu)}E_lB_k + \sum_{l,\,m,\,k=1}^{3}\sigma_{ilmk}^{(\nu)}E_lB_mB_k. \tag{50.5}$$

The rank II tensor was obtained previously in Sec. 40:

$$\sigma_{il}^{(\nu)} = \sigma_i^{(\nu)}\delta_{il} = \frac{e^2n_\nu(\tau)}{m_i}\,\delta_{il}. \tag{50.6}$$

The rank III tensor is of the form

$$\sigma_{ilk}^{(\nu)} = e^3n_\nu\frac{\langle\tau^2\rangle}{m_im_l}\,\varepsilon_{ilk}, \tag{50.7}$$

where ε_{ilk} is the so-called *unit antisymmetrical tensor* defined by the condition

$$\varepsilon_{ilk} = \begin{cases} \pm 1 \text{ for } i \neq l \neq k, \\ 0 \text{ for } i = l;\ l = k;\ k = i. \end{cases} \qquad (50.8)$$

The plus sign holds if *ilk* may be obtained from (1, 2, 3) as a result of even number of transpositions, and minus if the number of transpositions is odd, i.e.

$$\varepsilon_{123} = \varepsilon_{231} = \varepsilon_{312} = 1; \quad \varepsilon_{132} = \varepsilon_{321} = \varepsilon_{213} = -1, \qquad (50.9)$$

the other terms being zero. The expression (50.8) may be easily obtained by expanding the term $K_{12}^{(\nu)}[\mathbf{m}^{*} \mathbf{E}, \mathbf{B}]$. Find the *i*th component of $j_i^{(\nu)}$ determined by the second addend:

$$j_i^{(\nu)} = \sum_l K_{12}^{il(\nu)} [\mathbf{m}^{*-1}\mathbf{E},\ \mathbf{B}]_l. \qquad (50.10)$$

For example,

$$[\mathbf{m}^{*-1}\mathbf{E},\ \mathbf{B}]_x = \frac{E_y B_z}{m_y} - \frac{E_z B_y}{m_z} \qquad (50.11)$$

and

$$K_{12}^{ix(\nu)} = \frac{n_\nu \langle \tau^2 \rangle}{m_x} \delta_{ix}. \qquad (50.12)$$

We see that the sum (50.10) contains terms of the form $\frac{n_\nu \langle \tau^2 \rangle}{m_i m_l} \times$ $\times E_l B_k$ $(i \neq l \neq k)$, the sign of each term being dependent on the parity of the transpositions. Using the unit antisymmetrical tensor of rank III we arrive at the expressions (50.7) and (50.5). In the same way the third addend may be represented in the form (50.5), where $\sigma_{ijkl}^{(\nu)}$ may be written as follows:

$$\sigma_{ijkl}^{(\nu)} = n_\nu e^4 \langle \tau^3 \rangle \left[\frac{1}{3} \left(\frac{1}{m_1} + \frac{1}{m_2} + \frac{1}{m_3} \right) \right]^{-1} \times$$
$$\times \left[\frac{1}{3} \left(\frac{1}{m_1 m_2} + \frac{1}{m_2 m_3} + \frac{1}{m_3 m_1} \right) \right]^2 F_{ijkl}^{(\nu)} =$$
$$= n_\nu e^4 \langle \tau^3 \rangle \frac{(2m_t + m_l)^2}{3 m_t^3 m_l (m_t + 2m_l)} F_{ijkl}^{(\nu)}. \qquad (50.13)$$

The components of the *anisotropy tensor* $F_{ijkl}^{(\nu)}$ are written out in Table 16.

In order to obtain appropriate values of the components of the total tensors, the sum of corresponding ν-ellipsoid's tensor components, presented in the same co-ordinate system, should be calcu-

Table 16

Tensor components	Component value
$F^{(\nu)}_{1122}$	$-\frac{3(m_t+2m_l)m_t}{(m_l+2m_t)^2}$
$F^{(\nu)}_{1133}=F^{(\nu)}_{2233}$	$-\frac{3(m_t+2m_l)m_l}{(m_l+2m_t)^2}$
$F^{(\nu)}_{3311}=F^{(\nu)}_{3322}$	$-\frac{3(m_t+2m_l)m_l^2}{(m_l+2m_t)^2m_l}$
$F^{(\nu)}_{1212}=F^{(\nu)}_{1313}=F^{(\nu)}_{2323}$	$\frac{3(m_t+2m_l)m_t}{2(m_l+2m_t)^2}$
$F^{(\nu)}_{iiii}$	0

lated. Such calculations, if carried out, yield the following results: for conductivity

$$\sigma_{11}=\sigma_{22}=\sigma_{33}=\sigma_0=en\mu_d;$$
$$n=Mn_\nu;$$
$$\mu=\frac{e\langle\tau\rangle}{\tilde{m}^*};$$
$$\frac{1}{\tilde{m}^*}=\frac{1}{3}\left(\frac{2}{m_t}+\frac{1}{m_l}\right); \qquad (50.14)$$

for the Hall conductivity

$$\sigma_{ijk}=\frac{e^3n\langle\tau^2\rangle}{\tilde{\tilde{m}}^{*2}}\varepsilon_{ijk}; \quad \frac{1}{\tilde{\tilde{m}}^{*2}}=\frac{1}{3}\left(\frac{1}{m_l^2}+\frac{2}{m_lm_t}\right) \qquad (50.15)$$

and

$$\sigma_{ijkl}=\frac{\langle\tau^3\rangle}{\langle\tau\rangle^3}\sigma_0\mu^2\left(\frac{\tilde{m}}{\tilde{\tilde{m}}}\right)^4F_{ijkl}. \qquad (50.16)$$

The values of the F_{ijkl} coefficients are presented in Table 16 for two cases: the energy extrema are in the [100] direction, as in silicon, and in the [111] direction, as in germanium. The tensor σ_{ij} determines the current established by the electric field; σ_{ijk} determines the Hall current and voltage. The quantity σ_{ijkl} is connected with magnetoresistance. The galvanomagnetic effects may be described with the aid of the equation (50.5) by assigning appropriate values to the components of tensors of ranks II, III and IV. Figure 75 shows, by way of an example, the dependence of the magnetoresistance coefficient $H=\frac{1}{B^2}\left(\frac{\Delta\rho}{\rho_0}\right)$ on the angle between the current j and the magnetic field **B** for *n*-germanium.

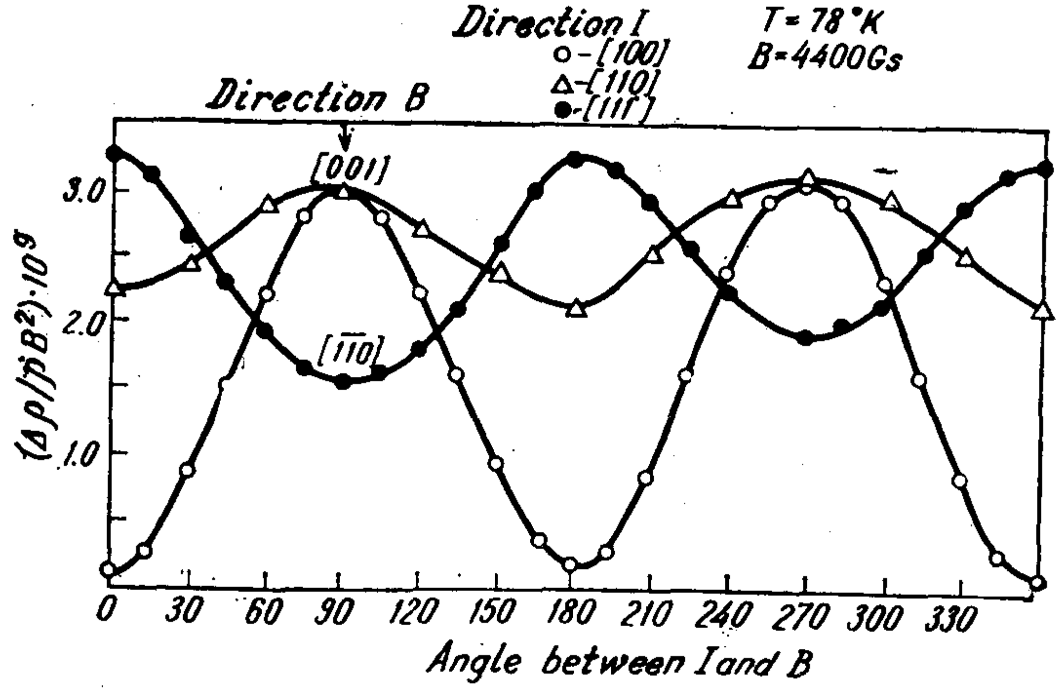


Fig. 75. The dependence of magnetoresistance on the angle between the current and the magnetic field

51. TENSORESISTIVE EFFECT. TENSOSENSITIVITY

Tensoresistive effect (or piezoresistance) is the name given to the variation of the resistance of a sample under stress. Tensoresistance does not follow directly from the kinetic equation. It must, however, be classified as a kinetic phenomenon since it represents the variation of one of the most important kinetic phenomena, the resistance, or conductivity, as a result of strains in the body produced by external loads.

To describe the tensoresistive effect we will need some results from the theory of elasticity, and they are summed up below.

In a deformed body the co-ordinates of its points are displaced. If the radius vector of a point in an undeformed body is **r**, and in the deformed body, **r**′,

$$\mathbf{u}=\mathbf{r}'-\mathbf{r} \tag{51.1}$$

is *the deformation*, or *displacement vector*. The state of the deformed body is fully described if **u** is determined as a function of the co-ordinates of the crystal points $\mathrm{u}(\mathbf{r})$. The deformation may be described with the aid of a symmetrical *strain tensor* u_{ik}:

$$\mathrm{u}_{ik}=\frac{1}{2}\left(\frac{\partial u_i}{\partial x_k}+\frac{\partial u_k}{\partial x_i}\right)=\mathrm{u}_{ki}. \qquad (51.2)$$

The strain tensor in its main axes representation is diagonal:

$$\mathrm{u}_{ik}=\mathrm{u}^{(i)}\delta_{ik}. \qquad (51.3)$$

The strain tensor u_{ik} *determines the variation of the distances between the points of a body upon deformation.*

If the distance between two points of an undeformed body is dl, and of the deformed body, dl', the relationship between dl' and dl may be expressed by the strain tensor. Indeed,

$$dl^2=dx^2+dy^2+dz^2=\sum_{i=1}^{3}dx_i^2 \qquad (51.4)$$

and

$$dl'^2=\sum_{i=1}^{3}dx'^2=\sum_{i=1}^{3}(dx_i+du_i)^2=\sum_{i=1}^{3}(dx_i^2+2dx_i\,du_i+du_i^2). \qquad (51.5)$$

Neglecting the quantity du_i^2 as being of the second order of magnitude and expressing du_i in terms of dx_k

$$du_i=\sum_{k=1}^{3}\frac{\partial u_i}{\partial x_k}dx_k, \qquad (51.6)$$

and taking into account (51.4) we obtain for dl'^2

$$dl'^2=\sum_{i=1}^{3}dx'^2=dl^2+2\sum_{i,\,k=1}^{3}\frac{\partial u_i}{\partial x_k}dx_i\,dx_k=$$

$$=dl^2+\sum_{i,\,k}\left(\frac{\partial u_i}{\partial x_k}+\frac{\partial u_k}{\partial x_i}\right)dx_idx_k=dl^2+2\sum_{i,\,k}\mathrm{u}_{ik}dx_i\,dx_k. \qquad (51.7)$$

If the strain tensor is of the diagonal form (51.3), (51.7) may be simplified:

$$dl'^2-dl^2=2\sum\mathrm{u}^{(i)}\delta_{ik}dx_i\,dx_k=2\sum_i\mathrm{u}^{(i)}dx_i^2. \qquad (51.8)$$

We may write from (51.8)

$$dl'=\sqrt{dl^2+2\sum_{i,k}\mathrm{u}_{ik}dx_i\,dx_k}=dl\sqrt{1+2\sum_{i,k}\mathrm{u}_{ik}n_in_k}. \qquad (51.9)$$

The relative elongation in the direction $\mathbf{n}=(n_1, n_2, n_3)$ *may be expressed with the aid of the strain tensor as follows:*

$$\mathrm{u}=\frac{dl'-dl}{dl}=\sqrt{1+2\sum_{i,k}\mathrm{u}_{ik}n_i n_k}-1\cong\sum_{i,k}\mathrm{u}_{ik}n_i n_k. \tag{51.10}$$

Find the variation of volume upon deformation:

$$\begin{aligned} d\tau' &= dx'\,dy'\,dz' = dx\,(1+\mathrm{u}^{(1')})\,dy\,(1+\mathrm{u}^{(2')})\,dz\,(1+\mathrm{u}^{(3')}) = \\ &= d\tau\,(1+\mathrm{u}^{(1')}+\mathrm{u}^{(2')}+\mathrm{u}^{(3')}), \end{aligned} \tag{51.11}$$

or

$$\frac{d\tau'-d\tau}{d\tau}=\Delta=\sum_l \mathrm{u}^{(l')}=\sum_l \mathrm{u}_{ll}=\sum_l \frac{\partial u_l}{\partial x_l}=\operatorname{div}\mathbf{u} \tag{51.12}$$

Thus, *the main values of the strain tensor determine the relative elongations along the tensor main axes, and their sum determines the relative variation of the volume.*

The strain tensor $\{\mathrm{u}_{ik}\}$ may be presented in a somewhat different form. To this end write an identity:

$$\mathrm{u}_{ik}=\left[\mathrm{u}_{ik}-\left(\frac{1}{3}\sum_{l=1}^{3}\mathrm{u}_{ll}\right)\delta_{ik}\right]+\left(\frac{1}{3}\sum_{l=1}^{3}\mathrm{u}_{ll}\right)\delta_{ik}. \tag{51.13}$$

The second tensor which is diagonal with identical diagonal elements, each equal to one third of the relative volume variation Δ, is termed *the cubic compression tensor.*

The tensor described by the term in the square brackets of (51.13) is *the shear tensor.* Since the sum of the diagonal elements of this tensor is zero, this means that *a pure shear is not accompanied by a variation of volume.*

Upon deformation internal stresses are created in the body which tend to return the body to the equilibrium undeformed state. They may be described with the aid of a rank II tensor, *the stress tensor* $\{\mathrm{p}_{ik}\}$. *The quantity* p_{ik} *is the projection of the force, acting on a plane of unit area perpendicular to the axis k, on the axis i.* The force $\mathbf{F}^{(1)}$ acting on a unit of the body volume may be expressed with the aid of the tensor $\mathbf{p}$ as follows:

$$\mathbf{F}^{(1)}=\operatorname{div}\mathbf{p}, \tag{51.14}$$

$$F_i^{(1)}=\sum_k \frac{\partial \mathrm{p}_{ik}}{\partial x_k}\quad (i=1,\,2,\,3). \tag{51.15}$$

There must be a connection between the strain and the stress tensors. Indeed, the stresses should increase with the strains. As long as deformations remain elastic, the dependence of the strains on the stresses, in compliance with Hooke's law, should be linear.

The quantity connecting stresses and strains is usually termed elasticity modulus. Since strains and stresses are generally tensors of rank II, *the elasticity modulus should be a tensor, too, and of a higher rank, namely rank* IV. Denote its elements by λ_{ijkl}. *The tensor* λ *is termed elasticity modulus tensor, or, simply, elasticity tensor.* Write in compliance with Hooke's law

$$\mathrm{p}_{ik}=\sum_{m}\lambda_{iklm}\mathrm{u}_{lm}. \qquad (51.16)$$

The elasticity tensor λ is symmetrical with respect to the pairs of its indices:

$$\lambda_{ijkl}=\lambda_{jikl}=\lambda_{ijlk}=\lambda_{klij}, \qquad (51.17)$$

since it connects two symmetrical tensors p_{ik} and u_{lm}. If (51.17) is taken into account it turns out that of the 81 tensor λ elements not more than 21 may be different. If, in addition, crystal symmetry is taken into account, the number of independent elasticity moduli may turn out to be still less. For instance, for the triclinic system the number of independent moduli is 18, for the rhombohedral—12, for the hexagonal—5, and for the cubic, only 3. The latter may be designated as follows:

$$\lambda_{xxxx};\quad \lambda_{xxyy};\quad \lambda_{xyxy}. \qquad (51.18)$$

Isotropic bodies are described by only two moduli, the shear and cubic compression moduli. For a body in a state of equilibrium the combined force acting upon a unit volume of the body is zero. Therefore, the condition of equilibrium is

$$F_i^{(1)}=\sum_{k}\frac{\partial \mathrm{p}_{ik}}{\partial x_k}=0. \qquad (51.19)$$

Suppose p is the external pressure acting upon the body. Take a surface element dS. The external force is counterbalanced by internal stresses:

$$p\,dS_i=p_i\,dS=\sum_{k}\mathrm{p}_{ik}\,dS_k. \qquad (51.20)$$

Denote $dS=\mathrm{n}\,dS$, where n is the external normal to the body surface. In this case

$$p_i\,dS=\sum_{k}\mathrm{p}_{ik}n_k\,dS, \qquad (51.21)$$

or

$$p_i=\sum_{k}\mathrm{p}_{ik}n_k, \qquad (51.22)$$

where n_k is the projection of the normal on the k-axis. For cubic compression $p_i = -p$. Since the pressure acts normally,

$$\mathrm{p}_{ik} = -p\delta_{ik}. \qquad (51.23)$$

If the body is *compressed unilaterally* $\mathbf{p} = (0, 0, p)$ then

$$\{\mathrm{p}_{ik}\} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -p \end{pmatrix}. \qquad (51.24)$$

For the case when the pressure is exerted along the direction $\mathbf{n} = (n_1, n_2, n_3)$ the stress tensor assumes the form

$$\mathrm{p}_{ik} = -pn_in_k. \qquad (51.25)$$

The connection (51.16) between p_{ik} and u_{lm} enables the strain to be expressed in terms of external pressure. As the number of independent components of the elasticity modulus tensor is far below 81, a two-index notation instead of the four-index one is used to describe it. For instance, for the cubic lattice crystals the following notation is used:

$$\lambda_{1111} \equiv \lambda_{xxxx} = \mathrm{c}_{11}; \quad \lambda_{1122} = c_{12}; \quad \lambda_{1212} = \mathrm{c}_{44}.$$

The c_{ik} components are symmetrical with respect to their indices.

Now let us describe the tensoresistive effect. Suppose the substance is described by the specific resistance tensor ρ^0 with the components ρ^0_{ik}.

If the semiconductor is strained its specific resistance will change to become ρ^t or ρ^t_{ik}. *The quantity* $\rho^t - \rho^0$ *or* $\rho^t_{ik} - \rho^0_{ik}$ *represents the variation of the specific resistance as the result of the load causing strains and stresses in the semiconductor*. The variation of the specific resistance may be expressed in two ways: either in terms of stresses, or in terms of strains. Since there is a definite relationship between the stress and the strain, both methods of description are equivalent. Express the variation of the resistance $\rho^t - \rho^0$ in terms of the stress tensor p_{ik} in the form

$$\rho^t_{ik} - \rho^0_{ik} = \rho^0_{ik} \sum_{lm} \pi_{iklm}\mathrm{p}_{lm}, \qquad (51.26)$$

or

$$\rho^t_{ik} = \rho^0_{ik}\left(1 + \sum_{lm} \pi_{iklm}\mathrm{p}_{lm}\right), \qquad (51.27)$$

$$\frac{\rho^t_{ik} - \rho^0_{ik}}{\rho^0_{ik}} = \sum_{lm} \pi_{iklm}\mathrm{p}_{lm}. \qquad (51.28)$$

The rank IV tensor π_{iklm} *is usually termed piezoresistance coefficients tensor, or piezoresistance tensor.*

For a cubic crystal only three of the coefficients of the piezoresistance tensor are different. For the sake of simplicity they are denoted by two-index letters:

$$\pi_{1111}=\pi_{11};\ \pi_{1122}=\pi_{12};\ \pi_{1212}=\pi_{44}. \qquad (51.29)$$

If (51.16) is taken into account ρ'_{ik} may be expressed in terms of the strain tensor and the elasticity modulus tensor in the form

$$\rho'_{ik}=\rho^0_{ik}\left(1+\sum_{lm}\pi_{iklm}p_{lm}\right)=\rho^0_{ik}\left(1+\sum_{lm,\ st}\pi_{iklm}\lambda_{lmst}u_{st}\right). \qquad (51.30)$$

For the experimental investigation of the tensoresistive effect the sample is chosen in the form of a rectangular cross section bar which is subjected to unilateral compression or extension. *If the current is directed along the strain axis, the effect will be longitudinal; if it makes an angle of 90° with the load axis, the effect will be transverse.*

Write the expression for the resistance ρ'_{ik} in the case of unilateral compression $(p>0)$ taking into accout (51.25):

$$\rho'_{ik}=\rho^0_{ik}\left(1+\sum_{lm}\pi_{iklm}p_{lm}\right)=\rho^0_{ik}\left(1-p\sum_{lm}\pi_{iklm}n_l n_m\right). \qquad (51.31)$$

Let the current $\mathbf{j}$ flow along the bar axis. Then $j_k=jn_k$. Using Ohm's law find the components of the field E_i:

$$\mathrm{E}_i=\sum_k\rho'_{ik}j_k=\sum_k\rho^0_{ik}\left(1-p\sum_{lm}\pi_{iklm}n_l n_m\right)jn_k. \qquad (51.32)$$

Denote the intensity of the field $\mathbf{E}$ along the bar axis by $\mathbf{E}^{\parallel}$:

$$\mathrm{E}^{\parallel}=(\mathbf{En})=\sum_{l=1}^{3}\mathrm{E}_l n_l. \qquad (51.33)$$

Apply the term *longitudinal specific resistance* $\rho^{\parallel}$ to the ratio of $\mathrm{E}^{\parallel}$ to j:

$$\rho^{\parallel}=\frac{E^{\parallel}}{j}=\frac{\sum_l \mathrm{E}_l n_l}{j}=\sum_{ik}\rho^0_{ik}\left(1-p\sum_{lm}\pi_{iklm}n_l n_m\right)n_k n_i. \qquad (51.34)$$

For $p=0$

$$\rho^{\parallel}=\sum_{ik}\rho^0_{ik}n_i n_k=\rho^{\parallel}_0. \qquad (51.35)$$

Longitudinal piezoresistance coefficient, or voltage piezosensitivity, are the terms applied to the quantity π_l (or π_{long})

$$\pi_l=\frac{\rho^{\parallel}-\rho^{\parallel}_0}{\rho^{\parallel}_0\,(-p)}. \qquad (51.36)$$

The piezoresistance along the sample axis may be expressed with the aid of the coefficient π_l:

$$\rho^l = \rho_0^l\,(1 - \pi_l p). \tag{51.37}$$

Express the longitudinal piezoresistance coefficient π_l *with the aid of the piezoresistance tensor using* (51.36), (51.35) *and* (51.34):

$$\pi_l = \sum_{iklm} \pi_{iklm} n_i n_k n_l n_m. \tag{51.38}$$

The sum (51.38) may be easily calculated if the use is made of (51.29):

$$\begin{aligned}\pi_l &= \pi_{11}\,(n_1^4 + n_2^4 + n_3^4) + 2\,(\pi_{12} + \pi_{44})\,(n_1^2 n_2^2 + n_1^2 n_3^2 + n_2^2 n_3^2) = \\ &= \pi_{11} + 2\,(\pi_{44} + \pi_{12} - \pi_{11})\,(n_1^2 n_2^2 + n_1^2 n_3^2 + n_2^2 n_3^2).\end{aligned} \tag{51.39}$$

If the sample (in the shape of a bar) is cut out so that its axis coincides with one of the [100] directions, only one of the **n** vector projections will be non-zero: $\mathbf{n} = (1, 0, 0)$, and therefore

$$\pi_{[100]} = \pi_x = \pi_y = \pi_z = \pi_{11}. \tag{51.40}$$

If the sample is cut along the [110] direction $\mathbf{n} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0\right)$, and

$$\pi_{[110]} = \frac{\pi_{11} + \pi_{12} + \pi_{44}}{2}. \tag{51.41}$$

In the same way we may obtain

$$\pi_{[111]} = \frac{2}{3}\,(\pi_{12} + \pi_{44}) + \frac{\pi_{11}}{3}. \tag{51.42}$$

Thus, *the voltage piezosensitivity coefficient* π_l *depends on the direction of the sample axis relative to the crystallographic axes in compliance with* (51.39).

One may introduce *the coefficient of piezosensitivity* S_l *to the strain* u along an axis.

The Young modulus E^Y, as is well known, is the quantity determined by the relation

$$-p = E^Y u, \tag{51.43}$$

where, according to (51.10),

$$u = \sum_{ik} u_{ik} n_i n_k. \tag{51.44}$$

Substituting (51.43) into (51.37) we obtain

$$\rho^l = \rho_0^l\,(1 - p\pi_l) = \rho_0^l\,(1 + E^Y \pi_l u) = \rho_0^l\,(1 + S_l u), \tag{51.45}$$

or

$$S_l = \frac{\rho^l - \rho_0^l}{u\rho_0} = \pi_l E^Y. \tag{51.46}$$

It must only be remembered that Young's modulus, as defined by (51.43), depends on direction. In compliance with (51.25) and (51.10) it may be expressed in terms of elasticity modulus as follows:

$$-p = \sum_{i,\,k} p_{ik} n_i n_k = E_n^Y \sum_{l,\,m} u_{lm} n_l n_m. \tag{51.47}$$

Expressing p_{ik} in terms of u_{lm} according to (51.16), we obtain

$$\frac{1}{E_n^Y} = \frac{u}{-p} = \frac{\sum\limits_{lm} u_{lm} n_l n_m}{\sum\limits_{iklm} \lambda_{iklm} u_{lm} n_i n_k}, \tag{51.48}$$

or

$$\frac{1}{E_n^Y} = \frac{c_{11} + c_{12}}{(c_{11} + 2c_{12})\,(c_{11} - c_{12})} + \left(\frac{1}{c_{44}} - \frac{2}{c_{11} - c_{12}}\right)(n_1^2 n_2^2 + n_1^2 n_3^2 + n_2^2 n_3^2). \tag{51.49}$$

For instance, for the [100] axis, $n = (1, 0, 0)$, we obtain

$$E_Y = \frac{c_{11} - c_{12}}{c_{11} + c_{22}}\,(c_{11} + 2c_{22}) = E_{[100]}^Y. \tag{51.50}$$

The piezosensitivity of semiconductors is several tens of times greater than that of metals. For instance, for *p*-silicon of 0.1 ohm·cm specific resistivity S_l amounts to about 125 (the relative strain piezosensitivity S_l is dimensionless). This is some 60 times greater than the corresponding figure for wire strain gauges.

52. PIEZORESISTIVE EFFECT. PIEZORESISTANCE COEFFICIENTS

Table 17 shows some experimental values of piezoresistance coefficients for germanium and silicon.

It may be seen from the table that both the piezoresistance coefficients π_l and the piezosensitivity S_l depend on the magnitude and the type of conductivity. They also depend on temperature and even on strain. To understand the dependences it is necessary to consider the physical nature of the piezoresistive effect. Generally, the specific resistivity tensor may be expressed in terms of the carrier concentration and the mobility tensor μ:

$$\rho^{-1} = en\mu = \sigma. \tag{52.1}$$

Table 17

Adiabatic Piezoresistance Coefficients

(T = 20°C)

Material	ρ_0, ohm·cm	π_{11}	π_{12}	π_{44}	$\pi_l = \pi_{[111]}$	$E^Y_{[111]}$, dyn/cm	$S_{[111]} = E^Y_{[111]} \times \pi_{[111]}$
		(10^{-12} cm²/dyn)					
Germanium:	1.5	—2.3	—3.2	—138.1	—94.9	1.55×10^{12}	—147
n-type	5.7	—2.7	—3.9	—136.8	—94.7		—147
	9.9	—4.7	—5.0	—137.9	—96.9		—150
	16.6	—5.2	—5.5	—138.7	—101.2		—157
p-type	1.1	—3.7	3.2	96.7	65.4		101.5
	15.0	10.6	5.0	46.5	31.4		48.7
Silicon:						1.87×10^{12}	
n-type	7.8	6.6	—1.1	138.1	93.6		175
p-type	11.7	—102.2	53.4	—13.6	—81.3		—142

The variation of ρ may be due both to the variation of the concentration and the mobility.

1. **Cubic compression.** The simplest case of piezoresistance is that of volume, or cubic, compression. The strain tensor reduces to a scalar $u_{ll} = u$. The lattice parameter a_0 of the crystal decreases:

$$a' = a_0(1 - u). \tag{52.2}$$

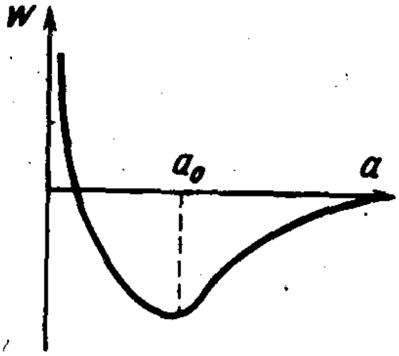


Fig. 76. The dependence of the energy of interaction between the atoms of the lattice on the distance between them

As the interatomic spacing decreases the overlapping of the electron wave functions becomes more pronounced, and this causes the potential energy $W(a)$, which describes the interaction of the lattice atoms, to change. If we accept the minimum of $W(a)$ to correspond to $a = a_0$, then $W(a)$ should increase both upon the compression and the extension of the crystal (Fig. 76). This variation of $W(a)$ and of the area of overlapping wave functions should lead to *the variation both of the exchange integral* $A(s)$ *and of the drop in the energy levels* C introduced in the theory *of the quasibound electron*. This, in turn, conditions *the variation of the forbidden and energy band widths.*

But the variation of the forbidden band width should cause the electron and hole concentrations to change. Denote by β *the coefficient of relative forbidden band width variation upon the varia-*

tion of pressure:

$$\beta = -\frac{1}{\Delta E_0}\frac{\partial \Delta E_0}{\partial p};$$
$$\Delta E_0(p) = \Delta E_0(1-\beta p). \tag{52.3}$$

Denoting the coefficient of relative forbidden band width variation upon deformation by γ,

$$\gamma = \frac{1}{\Delta E_0}\frac{\partial \Delta E_0}{\partial u},$$
$$\Delta E_0(\mathrm{u}) = \Delta E_0(1+\gamma \mathrm{u}), \tag{52.4}$$

and resorting to (52.4), (52.3) and (51.43), we may write:

$$\beta = \frac{\gamma}{E^Y}, \tag{52.5}$$

where E^Y is Young's modulus.

The variation of the forbidden band width is caused by the displacement of the bottom of the conduction band and the top of the valence band. It should, however, be kept in mind that *the displacement of E_c and E_v may not be the same.* Generally, the position of E_c and E_v is a function of the relative variation of the volume or of the strain tensor:

$$E_c(\mathrm{u}) = E_c + \frac{\partial E_c}{\partial \mathrm{u}}\mathrm{u} + \frac{1}{2}\frac{\partial^2 E_c}{\partial \mathrm{u}^2}\mathrm{u}^2 + \dots \tag{52.6}$$

Limiting the discussion to the case of small strains we may neglect all the terms of relative strain of orders above the first:

$$E_c(\mathrm{u}) = E_c(0) + \Delta_c \mathrm{u}. \tag{52.7}$$

The additional potential energy of the electron in a strained lattice is termed deformation potential, Δ_c being termed deformation potential constant. In the same way we may write for the top of the valence band

$$E_v(\mathrm{u}) = E_v(0) + \Delta_v \mathrm{u}. \tag{52.8}$$

The variation of the forbidden band width upon strain may be written in the form:

$$\Delta E_0(\mathrm{u}) = E_c(\mathrm{u}) - E_v(\mathrm{u}) = \Delta E(0) + (\Delta_c - \Delta_v)\,\mathrm{u}; \tag{52.9}$$

in other words, the quantity γ may be expressed in terms of Δ_c and Δ_v. Express the product of charge concentrations with the aid of the strain tensor:

$$np = n_i^2 = N_c N_v e^{-\frac{\Delta E_0(\mathrm{u})}{kT}} = n_i^2(0)\,e^{-\frac{(\Delta_c-\Delta_v)\,\mathrm{u}}{kT}} =$$
$$= n_i^2(0)\left(1 - \frac{\Delta_c - \Delta_v}{kT}\mathrm{u} + \dots\right). \tag{52.10}$$

Thus, *the increment of the product of electron and hole concentrations is related to the value of deformation potential constants and of the strain* **u**. In an intrinsic semiconductor $n=p$, and

$$\frac{\delta n}{n(0)}=\frac{\delta n_i}{n_i(0)}=\frac{\Delta_v-\Delta_c}{2kT}\mathbf{u}=\frac{\delta p}{p(0)}. \qquad (52.11)$$

Using the expression for conductivity we may write

$$\frac{\delta\rho}{\rho_0}=-\frac{\delta n}{n}=\frac{\Delta_c-\Delta_v}{2kT}\mathbf{u}=\frac{\Delta_v-\Delta_c}{2kT}\frac{p}{\mathrm{E}^Y}. \qquad (52.12)$$

Comparing (52.12) with (51.36) we may write the expression for the longitudinal piezoresistance coefficient:

$$\pi_l=\frac{\Delta_c-\Delta_v}{2kT\mathrm{E}^Y}; \quad S_l=\frac{\Delta_c-\Delta_v}{2kT}. \qquad (52.13)$$

The value of resistance variation upon compression obtained with π_l assessed on the basis of (52.13) is comparatively small. Since, according to (51.13), cubic compression generally constitutes a part of every strain except the pure shear, the variation of free carrier concentration caused by cubic compression should be observed in all cases.

However, in the impurity depletion range compression can change the total carrier concentration only by the amount equal to twice the minority carrier concentration, and for this reason piezoresistance will hardly be felt. In actual fact, however, piezoresistance often makes itself much more manifest, and this may be explained only on the basis of a complex energy-band pattern.

2. Unilateral extension or compression. Much greater values of piezoresistance are observed in cases of unilateral extension or compression when the stress tensor assumes the form (51.24) or (51.25). The strain is determined by the strain tensor $\mathbf{u}_{ik}$. The following point should be noted: when the interatomic spacing decreases in one direction, it increases in the perpendicular direction. The variation of atomic wave function overlapping will therefore be different for different directions.

Express the position of the bottom of the conduction band as follows:

$$E_c(\mathbf{u})=E_c(0)+\sum_{ik}\Delta_{ik}^{(c)}\mathrm{u}_{ik}, \qquad (52.14)$$

where $\mathbf{u}_{ik}$ is a strain tensor component, and $\Delta_{ik}^{(c)}$ deformation potential coefficients tensor component. The expression (52.14) is valid for small strains. Generalize the problem for the case of a M-valley semiconductor. We must presume that the position of the bottom of the conduction band in each valley (ν-valley) is described by its own tensor $\Delta_{ik}^{(c)(\nu)} = \Delta_{ik}^{(\nu)}$:

$$E_c^{(\nu)}(\mathbf{u}) = E_c(0) + \sum_{ik} \Delta_{ik}^{(\nu)} u_{ik}. \tag{52.15}$$

Since the strain tensor is the same for all valleys and $\Delta^{(\nu)}$ are, generally, different, *the displacement of the bottom of the conduction band will be different in each valley*. The position of the Fermi level is independent of the number of the valley, therefore the separation between the Fermi level and the conduction band bottom becomes different for each valley with the result that electron concentrations in them become different, too; if the bottom of the valley is raised, the number of electrons in it will decrease, if the bottom of the valley is lowered, the number of electrons in it will increase:

$$n^{(\nu)} = N_c e^{\frac{F' - E_c^{(\nu)}(\mathbf{u})}{kT}} = n^{(\nu)}(\mathbf{u}). \tag{52.16}$$

In our notation F' was the Fermi level in the strained crystal:

$$F' = F + \delta F; \quad \delta F = \sum_{ik} D_{ik} u_{ik}. \tag{52.17}$$

The case of extrinsic conductivity is the most interesting one. Considering, for the sake of simplicity, the impurity depletion range we may write the condition for the conservation of the number of conduction electrons:

$$\sum_{\nu=1}^{M} n^{(\nu)}(\mathbf{u}) = n^{(\nu)}(0)\, M = n. \tag{52.18}$$

Expanding (52.16) into a series of $\left(\delta F - \sum_{ik} \Delta_{ik} u_{ik}\right)\Big/ kT$ (which is small) and retaining only the linear term, we obtain

$$n^{(\nu)}(\mathbf{u}) = n^{(\nu)}(0)\left(1 - \frac{\delta F - \sum_{ik} \Delta_{ik} u_{ik}}{kT} + \ldots\right). \tag{52.19}$$

Substituting (52.19) into (52.18) we write

$$n^{(\nu)}(0)\, M = n^{(\nu)}(0)\, M - \frac{n^{(\nu)}(0)}{kT}\left[\sum_{\nu=1}^{M}\left(\delta F - \sum_{ik} \Delta_{ik}^{(\nu)} u_{ik}\right)\right]. \tag{52.20}$$

Equating the terms in square brackets (52.20) to zero and substituting (52.17) for δF we obtain

$$M\,\delta F = \sum_{\nu=1}^{M} \sum_{i,k=1}^{3} \Delta_{ik}^{(\nu)} u_{ik} = M \sum_{i,k=1}^{3} D_{ik} u_{ik}. \tag{52.21}$$

From (52.21) we have

$$D_{ik} = \frac{1}{M} \sum_{\nu=1}^{M} \Delta_{ik}^{(\nu)} = \langle \Delta \rangle_{ik}, \qquad (52.22)$$

i.e. *the tensor D_{ik}, which determines the Fermi level displacement required to keep the number of conduction electrons constant, is the tensor $\Delta_{ik}^{(\nu)}$ averaged over the minima.* Taking into account (52.22) re-write the expression for $n^{(\nu)}$ (u):

$$n^{(\nu)}(\mathbf{u}) = n^{(\nu)}(0) \left[1 - \frac{\sum_{ik} (\langle \Delta \rangle_{ik} - \Delta_{ik}^{(\nu)}) \, u_{ik}}{kT} \right] = n^{(\nu)}(0) + \delta n^{(\nu)}. \qquad (52.23)$$

Since the electron concentration at different minima varies differently upon strain, *the minima cease to be equivalent.* Write now the expression for full conductivity:

$$\sigma(\mathbf{u}) = \sum_{\nu=1}^{M} \sigma^{(\nu)}(\mathbf{u}) = e \sum_{\nu=1}^{M} n^{(\nu)}(\mathbf{u}) \, \mu^{(\nu)}. \qquad (52.24)$$

If the constant-energy surfaces at the minima are spherical, $\mu^{(\nu)} = \mu$ will not depend on the valley number, and according to (52.18) $\sigma(\mathbf{u}) = \sigma(0)$, i.e. there will be no piezoresistance in this case. *If, however, $\mu^{(\nu)}$ is anisotropic, the conductivity will become anisotropic upon strain:*

$$\sigma(\mathbf{u}) = \sigma(0) - e \sum_{\nu=1}^{M} \mu^{(\nu)} \delta n^{(\nu)}. \qquad (52.25)$$

For a fixed strain corresponding to definite $\delta n^{(\nu)}$ *the conductivity anisotropy is related to the mobility anisotropy, i.e. to the anisotropy of the inverse effective mass.* Consider *n*-silicon as an example.

Should a silicon crystal be stressed in the [100] direction, the interatomic spacing in this direction would decrease, the value of the exchange integral for this direction would grow, falling at the same time for the directions [001] and [010]. The bottom of the conduction band would drop in the [100] direction and rise in the perpendicular directions.

Since

$$\sum_{\nu=1}^{M} \delta n^{(\nu)} = 0; \qquad (52.26)$$

the area of occupied states at the minimum ($\pm k_{0x}$, 0, 0) should grow, decreasing at minima of other types (Fig. 77). Denote the increase in electron concentrations at the first and the fourth minima by δn_1. Evidently,

$$-4\delta n_2 = 2\delta n_1. \tag{52.27}$$

Write the variation of the conductivity tensor $\delta\sigma$:

$$\delta\sigma_{xx} = -e2\delta n_1 (\mu_{xx}^{(1)} - 2\mu_{xx}^{(2)}) = -e^2\langle\tau\rangle 2\delta n_1 \left(\frac{1}{m_l} - \frac{2}{m_t}\right), \tag{52.28}$$

$$\delta\sigma_{yy} = -e^2\langle\tau\rangle \left(\frac{2\delta n_1}{m_t} + \frac{2\delta n_2}{m_t} + \frac{2\delta n_3}{m_l}\right) = -e^2\langle\tau\rangle \delta n_1 \left(\frac{1}{m_t} - \frac{1}{m_l}\right) = \delta\sigma_{zz}. \tag{52.29}$$

Thus, both the sign and the modulus of conductivity variation are different for different crystallographic axes. If the pressure is applied along the [110] axis, the minima *1*, *2*, *4*, and *5* will drop, and *3* and *6* will rise; $4\delta n_1 = -2\delta n_3$. The expression for conductivity variation will assume the form

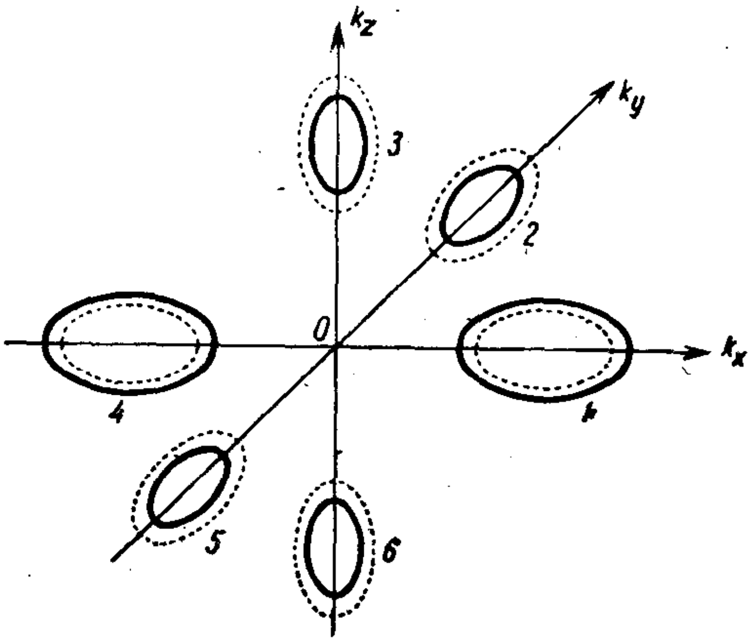


Fig. 77. The redistribution of the electrons among the energy minima upon deformation in silicon

$$\delta\sigma_{xx} = -2e^2\delta n_1 \langle\tau\rangle \left(\frac{1}{m_l} - \frac{1}{m_t}\right) = \delta\sigma_{yy}, \tag{52.30}$$

$$\delta\sigma_{zz} = -4e^2\langle\tau\rangle \delta n_1 \left(\frac{1}{m_t} - \frac{1}{m_l}\right). \tag{52.31}$$

Should the crystal be compressed in the [111] direction, all the extrema would remain equivalent; there would be no redistribution of electrons, and no piezoresistance should be observed in this approximation. Actually it will be in evidence because of the charge carrier concentration variation mentioned in the first item.

Consider now piezoresistance of p-germanium and p-silicon. The variation of resistivity in these semiconductors is so great that it cannot be explained by the above-mentioned variation of carrier concentration.

Recall that the constant-energy surfaces of the valence band in germanium and silicon are almost spherical, therefore the anomaly of a large piezoresistance cannot be explained, as in the case of the conduction band, by the anisotropy of conductivities. The explanation found was based on the degeneracy of the valence band: two energy branches, for the light and heavy holes, merge at the point $\mathbf{k}=0$. *When anisotropic strain is applied the lattice field symmetry is disturbed, and as a result the degeneracy is removed since the top of the valence band for the light and the heavy holes is displaced by a different amount in opposite directions.* The displacement of the energy levels $\delta E(\mathbf{k})$ depends on $\mathbf{k}$ in the following way:

$$\delta E(\mathbf{k}) = \pm \frac{bB}{2\bar{\mathrm{B}}}\left(\frac{3}{k^2}\sum_{ij} \mathrm{u}_{ij}k_ik_j - \sum_{i} \mathrm{u}_{ii}\right), \qquad (52.32)$$

where $\bar{\mathrm{B}} = \left(B^2 + \frac{1}{5}C^2\right)^{1/2}$; B, C are constants in the expression (26.2); b is the deformation potential constant. The upper sign is for the light holes, the lower, for the heavy. The displacement of the energy bands of the light and heavy holes changes their concentration, at the same time retaining the total number of conduction holes (in case of extrinsic conductivity in the impurity depletion range): $p_l + p_h = N_a$. But the redistribution of the concentrations of light and heavy holes results in the change in the conductivity because of the difference in their mobilities:

$$\delta\sigma = e_p(\delta p_l \mu_{pl} + \delta p_h \mu_{ph}) \qquad (52.33)$$

for $\delta p_l = -\delta p_h$.

Thus, the anomaly of a large piezoresistance in p-germanium and p-silicon is related to the difference in effective masses of light and heavy holes and, as a consequence, in their mobilities.

Should a strained crystal be placed in a magnetic field, all galvano- and thermomagnetic phenomena should take a different course than in an undeformed crystal. The changes in all the kinetic phenomena are due to the fact that crystal strain occasions changes in its energy-band pattern. *For a more accurate description of phenomena taking place in strained semiconductors the changes in the scattering mechanism, in the effective mass and in other parameters should be taken into account.*

The expression of the type (50.5) for current density will be changed when the crystal is subjected to strain. Using the Ohm law we may write the expression for the current density in the form

$$j_i = \sum_{k=1}^{3} \sigma_{ik}(\mathbf{u},\ \mathbf{B})\, E_k, \tag{52.34}$$

where the conductivity σ_{ik} may be expressed for the case of weak fields in the form

$$\sigma_{ik}(\mathbf{u},\mathbf{B}) = \sigma^0_{ik}\left[1 + \sum_{lm}(\pi^{-1})_{iklm}\mathrm{u}_{lm}\right] + \left\{\left[\sum_{l}\sigma_{ikl} + \sum_{lmn}\sigma_{iklmn}\mathrm{u}_{mn}\right]B_l\right\} + \\ + \left\{\left[\sum_{lm}\sigma_{iklm} + \sum_{lmnp}\sigma_{iklmnp}\mathrm{u}_{np}\right]B_lB_m\right\}. \tag{52.35}$$

In the absence of strains ($\mathrm{u}_{st}=0$), (52.35) will be reduced to the expression of type (50.5); for $\mathbf{B}=0$ it will be reduced to (51.30), if it is taken into account that the tensors ρ_{ik} and σ_{ik} are reciprocal, and $(\pi^{-1})_{iklm}$ denotes the *iklm* element of the tensor reciprocal of the piezoresistance tensor. The tensors σ_{ikl} and σ_{iklm} describe the Hall conductivity and the magnetoresistive effect. The tensors σ_{iklmn} and σ_{iklmnp} describe the transverse and the longitudinal piezogalvanomagnetic effects, respectively.

For a semiconductor with M extrema the sum of the expression (52.35) over all the extrema should be taken.

Summary of Secs. 51-52

1. The piezoresistive effect consists in the variation of the resistivity (or conductivity) of a semiconductor or a metal, caused by strain. The physical cause of piezoresistance is the variation of the semiconductor energy band pattern. The variation of the forbidden band width occasions the variation in the charge carrier concentration and, consequently, in the resistivity.

2. In substances with complex band patterns, such as n-germanium and n-silicon, crystal strain caused by unilateral stress or strain results in large resistivity variations which cannot be explained by the variation of the total carrier concentration. The explanation rests on the fact that the energy extrema become unequivalent as a result of anisotropic strain, and that a re-distribution of electrons over the extrema takes place. The contribution to conductivity of the minima whose bottoms sink proves to be greater than that of the minima the bottoms of which rise. In such cases the conductivity variation may be observed only for non-spherical energy surfaces.

3. Large resistivity variations in semiconductors of the p-silicon type are explained by the removal of the energy band degeneracy

resulting from the application of anisotropic strain. Since the mobilities of light and heavy holes are different their contribution to conductivity is different, too, and this results in variations of resistance even if the total number of holes remains constant.

4. The piezoresistive effect is described by the tensor of piezoresistance coefficients. This is a rank IV tensor.

If a magnetic field and strain are simultaneously applied to a semiconductor, its conductivity may be represented, in the case of weak fields, in the form (52.35).

Chapter V

THE THEORY OF CHARGE CARRIERS SCATTERING

53. EFFECTIVE SCATTERING CROSS SECTION

In an ideal crystal the directional motion of electrons and holes can continue for an infinitely long time even in the absence of external electric field.

In the real crystal, after the field has been switched off, the electric current falls off exponentially:

$$\mathbf{j}(t) = \mathbf{j}_0 e^{-\frac{t}{\langle\tau\rangle}} \tag{53.1}$$

where $\langle\tau\rangle$ is a parameter of the crystal of the order of 10^{-13} s, i.e. the current drops to zero practically instantly. *The mechanism which returns the system to the equilibrium state is the charge carrier scattering, their collisions with lattice periodic field imperfections of various types.*

The scattering process is characterized by a quantity usually termed *effective cross section*. It may be introduced as follows.

Suppose one particle flows through a unit cross section per unit time with the velocity $\mathbf{v}$; it constitutes a *single-particle flow*. Defining the flow as the number n of particles passing through a unit cross section per unit time we may express n in terms of the concentration of particles in the stream n_1 and their velocity v: $n = n_1 v$. For $n = 1$ cm^{-2}s^{-1} the concentration of particles in the stream is $n_1 = \frac{1}{v}$ cm^{-2}s^{-1}/(cm·s^{-1}) $= \frac{1}{v}$ cm^{-3}. Place a screen impenetrable to the particles in some area of the unit cross section. The probability of a particle colliding with the screen is $S: 1 = S$, where S is the area of the screen, provided the particle is sure to pass through the unit cross section (1 particle per 1 cm^2 per 1 s). If each collision with the area S results in the particle leaving the stream then S will be equal to *the probability of collision, or of scattering*. Thus, knowing the area of the screen, we know the probability of the particle colliding with the screen.

If the number of screens per unit cross section is N, the probability will be N times greater, i.e. NS. If the number of particles passing through unit cross section per unit time is not one, but n, the number of particles scattered per unit time will be $\Delta n = nNS$.

When particles belonging to a stream of unit cross section move in a medium containing scattering centres with a concentration N, each section of their path dx long will contain $N \cdot 1 \cdot dx$ such centres, and the probability for a particle to be scattered by them will be $SNdx$. The number of particles scattered per unit time will be $dn = nNSdx$. If the number of particles in the stream is $n(x)$ at the point x, at the point $x+dx$ it will be smaller by the amount $dn = nSNdx$ so that $n(x+dx) = n(x) + dn$. Since the particle flow decreases, $n(x+dx) < n(x)$ for $dx > 0$ and, therefore, $dn < 0$:

$$-dn = n(x)\,SNdx. \qquad (53.2)$$

The equation (53.2) may be easily solved if S and N are independent of x:

$$n(x) = n_0 e^{-SNx}. \qquad (53.3)$$

We see from here that knowing the scattering probability S we can calculate particle flow at any point x. The number of particles in the flow, or the flow itself, decreases exponentially. *At a distance of $\Delta x = \frac{1}{NS}$ the number of particles decreases e times.* A somewhat different meaning may be attributed to the quantity $\frac{1}{NS}$. To this end take an interval x, $x+dx$, $-dn(x)$ particles being scattered inside this interval. Suppose all the particles are travelling from the point $x=0$. All the particles that are scattered inside the interval x, $x+dx$ have travelled a distance, from $x=0$ to $x+dx$, or within an error dx, the distance x. The total distance covered by such particles is

$$-dn(x)\,x. \qquad (53.4)$$

If we perform summation over all the values of x we shall obtain the sum of sections travelled by all the n_0 particles without scattering:

$$l = \frac{1}{n_0}\int_0^\infty x(-dn) = -\frac{n_0}{n_0}\int_0^\infty xe^{-SNx}SN\,dx =$$

$$= \frac{-1}{SN}\int_0^\infty ye^{-y}\,dy = \frac{1}{SN}. \qquad (53.5)$$

The expression (53.5) connects the mean free path l with the effective scattering cross section. But it follows from (53.2) that the scattering probability over a section of the path dx long is

$$SN\,dx=\frac{dx}{l}. \tag{53.6}$$

Consequently, *the quantity $SN=l^{-1}$ is the scattering probability per unit length of the path.* This relationship is quite analogous to the one between the mean free time and the probability of scattering per unit time obtained in Sec. 2 with the aid of the theory of probability. A relation between these quantities may be established. The modulus of the mean velocity v may be defined by the relation

$$v=\frac{l}{\tau}. \tag{53.7}$$

In this case

$$\tau=\frac{l}{v}=\frac{1}{vSN}. \tag{53.8}$$

The quantity S was introduced as the area of the "screen" on which the particles belonging to the stream were scattered. To describe scattering processes of electrons moving in a solid several more complex concepts should be introduced. *We will apply the term scattering centre to any imperfection of the ideal lattice field occasioning a change in the quasimomentum and the velocity of the electron, or hole.* The interaction of the charge carriers with these centres is described by some probability W.

Build a screen around the scattering centre with an area $S^*=W$. The probabilities of scattering by the screen and by the scattering centre will in this case be equal, for this reason *scattering probability is often termed effective scattering cross section.*

If several scattering processes are possible each described by an effective cross section S_i^* and by a concentration of scattering centres N_i of type i, the combined scattering probability W will be equal to the sum of individual probabilities:

$$W=S^*=\sum_i N_iS_i^*. \tag{53.9}$$

The quantity W may be described by the mean free path $l=\frac{1}{W}$. At the same time $S_iN_i=\frac{1}{l_i}$, and

$$\frac{1}{l}=\sum_i\frac{1}{l_i}. \tag{53.10}$$

In other words, if several scattering mechanisms are active, and the i-th mechanism acting individually results in a free mean path l_i, the simultaneous action of all these mechanisms will result in a mean free path l which, naturally, will be less than any of the individual mean free paths, and this is exactly what the relation (53.10) states. The relation expresses the property of additivity of probabilities of independent events. It follows from (53.10) that the resultant mean free path is always smaller than the smallest partial mean free path. To assess the contribution of some scattering mechanism to the combined mean free path l, information on the concentration of the corresponding scattering centres and on their effective scattering cross section should be available.

The effective cross section S^* may be a function of the energy, effective mass, or of some other parameter of the particles being scattered. The task of the theory of scattering is to establish the dependence of S^* on the energy and effective mass of the particle for various types of scattering centres.

We defined the effective cross section S^* as the probability for a particle to be scattered, i.e. to be deflected through an arbitrary angle from its initial direction of motion. However, as the scattering process is a random process, various particles will be deflected through various angles θ, φ from the direction of their motion. Suppose dn' particles are deflected into a solid angle element $d\Omega = \sin\theta d\theta d\varphi$ over an interval dx per unit time. This number may be expressed in the form

$$dn' = dn'(\theta, \varphi) = nN\, dx\, \sigma(\theta, \varphi)\, d\Omega, \qquad (53.11)$$

whence

$$\sigma(\theta, \varphi) = \frac{dn'}{nN\, dx\, d\Omega}. \qquad (53.12)$$

The quantity $\sigma(\theta, \varphi)$ is numerically equal to the number of particles dn' scattered per unit time into a unit solid angle $d\Omega = 1$ over a section of the path $dx = 1$ by one scattering centre $N = 1$ in case of a unit flow $n = n_1 v = 1$. The quantity

$$\sigma(\theta, \varphi)\, d\Omega = \frac{dn'}{1} \quad (n_1 v = n = 1; \quad N\, dx = 1) \qquad (53.13)$$

represents the scattering probability of a particle belonging to a single particle flow by one scattering centre through the angles θ and φ into the solid angle $d\Omega$. For this reason $\sigma(\theta, \varphi)$ is termed differential effective cross section. The dimensionality of $\sigma(\theta, \varphi)$ may be found from (53.13). Since dn' is the scattering probability for one particle per unit time ($[dn'] = [T]^{-1}$) for a single-particle flow ($n_1 v = 1$ cm^{-2} s^{-1}, i.e. $[n_1 v] = [L]^{-2}[T]^{-1}$) it follows that

$$[\sigma(\theta, \varphi)] = \frac{[T]^{-1}}{[L]^{-2}[T]^{-1}} = [L]^2. \qquad (53.14)$$

We see from here that the differential effective cross section dimensionality is that of area, the same as of the effective cross section S^*.

Performing the summation of the quantities in (53.13) over all angles θ, φ we obtain the probability of scattering through an arbitrary angle, i.e. the combined scattering probability S^*:

$$\int_{(4\pi)} \sigma(\theta, \varphi)\, d\Omega = S^*. \tag{53.15}$$

The equation (53.15) represents the condition of normalizing $\sigma(\theta, \varphi)$. The quanity S^* may be termed *integral effective cross section*. If the differential effective cross section is independent of the angles θ and φ, i.e. the probability of scattering through any angle is the same, such scattering is termed *isotropic*. Obviously, for the case of isotropic scattering the relation between σ and S^* will be simple:

$$S^* = \int_{(4\pi)} \sigma(\theta, \varphi)\, d\Omega = \sigma \int_{(4\pi)} d\Omega = 4\pi\sigma, \tag{53.16}$$

$$\sigma = \frac{S^*}{4\pi}. \tag{53.17}$$

The angle θ is the angle between the initial direction of motion of the particle and its direction of motion after scattering. The angle φ is measured from some plane passing through the polar axis (the axis Oz) coinciding with the direction of initial motion of the particle. If the scattering centre is axially symmetrical, σ will depend only on the polar angle θ and will be equal to the probability of scattering into a solid angle contained between two cones $\theta + d\theta$ and θ (Fig. 78).

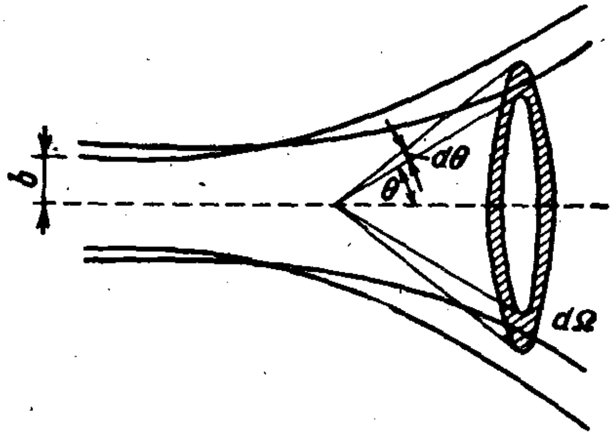


Fig. 78. The conical space angle of scattering

Assuming σ to be independent of φ and integrating (53.13) over φ, we obtain

$$\sigma(\theta) \sin\theta\, d\theta \int_0^{2\pi} d\varphi = \sigma(\theta)\, 2\pi \sin\theta\, d\theta = \sigma(\theta)\, d\Omega, \tag{53.18}$$

where $d\Omega = 2\pi \sin\theta\, d\theta$ is the angle between the two cones.

The collisions, as is well known, may be elastic or inelastic. *Collisions are termed elastic if the kinetic energy of the colliding particles is conserved in the process.* If, on the other hand, the kinetic energy of the particles after the collision is greater (or less) than it was before the act, the collision is termed *inelastic.*

When a light particle collides with a heavy immobile scattering centre its energy in case of elastic scattering remains practically the same. To test this statement write the laws of conservation of energy and momentum:

$$\mathbf{p}_1 + 0 = \mathbf{p}_2 + \mathbf{P}, \tag{53.19}$$

$$T_1 + 0 = \frac{p_1^2}{2m} + 0 = \frac{p_2^2}{2m} + \frac{p^2}{2M} = T_2 + T, \tag{53.20}$$

where $(0, \mathbf{P})$ and $(\mathbf{p}_1, \mathbf{p}_2)$ are the momenta of the heavy and light particles before and after the collision, respectively.

Expressing $\mathbf{P}$ from (53.19) and substituting into (53.20) we obtain

$$\mathbf{P} = \mathbf{p}_1 - \mathbf{p}_2; \quad P^2 = p_1^2 + p_2^2 - 2p_1p_2\cos\theta. \tag{53.21}$$

We may write from (53.20)

$$P^2 = \frac{M}{m}(p_1^2 - p_2^2). \tag{53.22}$$

Comparing (53.22) with (53.21) we obtain

$$p_1^2 + p_2^2 - 2p_1p_2\cos\theta - \frac{M}{m}p_1^2 + \frac{M}{m}p_2^2 = 0, \tag{53.23}$$

or

$$p_2^2 - \frac{2p_1\cos\theta}{\left(1+\frac{M}{m}\right)}p_2 + \frac{1-\frac{M}{m}}{1+\frac{M}{m}}p_1^2 = 0. \tag{53.24}$$

Solve the equation (53.24)

$$p_2 = \frac{p_1\cos\theta}{1+\frac{M}{m}} \pm \sqrt{\frac{p_1^2\cos^2\theta}{\left(1+\frac{M}{m}\right)^2} - \frac{\left(1-\frac{M}{m}\right)}{\left(1+\frac{M}{m}\right)}p_1^2} =$$

$$= p_1\frac{\cos\theta \pm \sqrt{\frac{M^2}{m^2} - \sin^2\theta}}{1+\frac{M}{m}}. \tag{53.25}$$

For the case $\frac{m}{M} \ll 1$ we retain in (53.25) only the terms of the first order of magnitude of $\frac{m}{M}$, omit the minus sign before the radical

$(p_2 > 0)$ and obtain

$$p_2 = p_1 \frac{\cos\theta + \frac{M}{m}}{1 + \frac{M}{m}} = p_1 \left[1 - \frac{m}{M} (1 - \cos\theta) \right]. \tag{53.26}$$

We see from here that the modulus of the momentum of the scattered particle is somewhat smaller than that of the incident particle:

$$p_1 - p_2 = p_1 \frac{m}{M} (1 - \cos\theta). \tag{53.27}$$

Find the variation of the energy of the scattered particle:

$$\Delta T = T_1 - T_2 = T = \frac{1}{2m} (p_1^2 - p_2^2) = \frac{1}{2m} (p_1 + p_2)(p_1 - p_2) =$$
$$= \frac{1}{2m} 2p_1 \frac{p_1 m}{M} (1 - \cos\theta) = \frac{p_1^2}{2m} \frac{2m}{M} (1 - \cos\theta) =$$
$$= T_1 \frac{2m}{M} (1 - \cos\theta). \tag{53.28}$$

The fraction of the energy transmitted by the incident particle to the scattering centre depends on the scattering angle: $\Delta T = T = 0$ for $\theta = 0$ (there is no scattering). The maximum variation of the incident particle energy takes place when the scattering angle is $\theta = \pi$, i.e. when the particle rebounds backwards:

$$\Delta T = T = T_1 \frac{4m}{M}. \tag{53.29}$$

Find the average energy transmitted by the scattered particle to the scattering centre in one collision. To this end take the probability of scattering into the angular interval $d\theta$ to be, in compliance with (53.18), equal to $\sigma(\theta)\, 2\pi \sin\theta\, d\theta$; this, at the same time, is the probability for the particle to loose the energy $\Delta T(\theta)$. Multiplying $\Delta T(\theta)$ by the probability of scattering and taking into account that the function $\sigma(\theta)$ is normalized to S^*, we write

$$\langle \Delta T \rangle = \frac{\int_0^\pi \Delta T(\theta)\, \sigma(\theta)\, 2\pi \sin\theta\, d\theta}{\int_0^\pi \sigma(\theta)\, 2\pi \sin\theta\, d\theta} = T_1 \frac{2m}{M} \frac{2\pi \int_0^\pi \sigma(\theta)(1 - \cos\theta) \sin\theta\, d\theta}{S^*} =$$
$$= T_1 \frac{2m}{M} \frac{\sigma_c}{S^*}, \tag{53.30}$$

where

$$\sigma_c = 2\pi \int_0^\pi \sigma(\theta)(1 - \cos\theta) \sin\theta\, d\theta \tag{53.31}$$

is the *averaged differential effective cross section*; the averaging over the angles θ is performed with the aid of the weighting function $(1-\cos\theta)$; σ_c *is termed effective conductivity, or mobility, or transport, cross section*. The weighting function decreases the contribution of $\sigma(\theta)$ to σ_c for small-angle scattering and increases almost twofold the contribution of large-angle scattering. This is quite understandable, for the fraction of energy transmitted is larger for large scattering angles. Consider the simplest case of isotropic scattering. Calculate the effective transport cross section for isotropic scattering:

$$\sigma_c = 2\pi\sigma \int_0^\pi (1-\cos\theta)\sin\theta\, d\theta = 4\pi\sigma = S^*. \tag{53.32}$$

Hence, *for isotropic scattering* $\langle\Delta T\rangle = \frac{2m}{M} T_1$.

The quantity σ_c may be interpreted in a somewhat different way. Since the scattering through an angle θ leads to variations of the velocity and the momentum connected with the directional motion, we may find the fraction of the directional speed v_z lost as the result of the collision. Taking into account that

$$v_{z2} \cong v_{z1}\cos\theta, \tag{53.33}$$

we may write

$$\Delta v_z = v_1 - v_{z2} \cong v_z(1-\cos\theta), \tag{53.34}$$

or

$$\frac{\Delta v_z}{v_z} = 1-\cos\theta. \tag{53.35}$$

Averaging the quantity Δv_z over the angles θ we obtain

$$\langle\Delta v_z\rangle = v_z\langle 1-\cos\theta\rangle = \frac{v_z}{S^*}2\pi\int_0^\pi (1-\cos\theta)\,\sigma(\theta)\sin\theta\, d\theta = v_z\frac{\sigma_c}{S^*}. \tag{53.36}$$

The ratio $\frac{\langle\Delta v_z\rangle}{v_z}$ is equal to the ratio $\frac{\sigma_c}{S^*}$. If we denote $\frac{\sigma_c}{S^*}$ by q^{-1}, then

$$\langle\Delta v_z\rangle = \frac{v_z}{q}; \quad q = \frac{v_z}{\langle\Delta v_z\rangle} = \frac{S^*}{\sigma_c}. \tag{53.37}$$

The quantity q may be regarded as *the number of collisions resulting in the complete loss of the velocity of directional motion*. Since for isotropic scattering $\sigma_c = S^*$, it follows that *in case of isotropic scattering the velocity of directional motion is completely lost in one collision*.

In conclusion of the section we present a list of the main types of scattering centres:

1) impurity ions;
2) thermal lattice vibrations;
3) impurity atoms;
4) vacancies and point defects;
5) dislocations;
6) grain boundaries, cleavage planes, crystal surfaces;
7) charge carriers.

The parts played by the different scattering centres will, evidently, be different. To make an assessment one should find the scattering probability for each of the scattering centre types listed above. To this end it suffices to evaluate an area within which the interaction between the charge carrier and the scattering centre is possible.

The mutual scattering of charge carriers should not be of much importance since their interaction was already accounted for by the introduction of the self-consistent field. This field enabled the charge carriers to be considered as non-interacting particles. This, naturally, does not mean that electrons do not conform to Coulomb's law. The interaction obeys the usual laws of electrodynamics, but because of the large numbers of interacting particles and of their wave properties, their real motion proceeds as if there would practically be no interaction, i.e. *the motion itself is in a sort of dynamic stability* contained in the Bloch functions and *the quasimomentum conservation law*. A conclusion may be drawn from the above that the scattering on charge carriers cannot be the *principal* mechanism which determines the course of kinetic phenomena. The scattering on the crystal surface, on the grain boundaries should play an important part in mono- or polycrystalline films. In the volume of a *single crystal* the part played by this factor is of no importance.

Dislocations penetrate large areas of the crystal, therefore their effective cross section should be very great. If the linear dislocation is 1 mm long and its diameter is of the order of a hundred lattice parameters the area of its axial cross section will be equal to $5\times$ $\times 10^{-8}\times 10^{2}\times 10^{-1}=5\times 10^{-7}$ cm^{2}. If the volume density of dislocations is $N_D \cong 10^{8}$ cm^{-3}, $l_D = S_D^{*-1} N_D^{-1} = 2\times 10^{-2}$ cm. As will become clear from the estimates presented below this value is rather high, and some phenomena may not be observed in crystals with such high dislocation densities.

For point defects and impurity atoms S^* may be assumed to be equal to an area of the order of the elementary cell face, i.e. $S_a^* =$ $= 3\times 10^{-15}$ cm^{2}. For $N_a \cong 10^{16}$ cm^{-3}, $l_a \cong 3\times 10^{-2}$ cm.

For impurity atoms S_l^* should, probably, be assumed to be larger than S_a^* by at least two orders of magnitude, and, accordingly, we obtain for $N_l \cong 10^{16}$ cm^{-3} and $S_l^* \cong 3\times 10^{-13}$ cm^{2}, $l_l \cong 3\times$ $\times 10^{-4}$ cm $= 3$ μm. When evaluating the part played by the thermal vibrations of the lattice one should keep in mind that S_T^* should be taken equal not to the cross section of the matrix atom, but to *the cross section of the area it occupies when taking part in thermal vibrations*, i.e. the amplitude of vibrations should be taken into account. Obviously, the amplitude of vibrations should be the greater, the higher the temperature. If it is assumed that the amplitude of vibration is of the order of 1 Å, the cross section of the area occupied by the atom in the course of vibrations (without the area occupied by the stationary atom) will be equal to twice the product of the atom diameter by the vibration amplitude, i.e. $S_T^* \cong 10^{-16}$ cm². We see from here that the effective scattering cross section of thermal vibrations is less than of all the other types of scattering centres considered above. Since, however, the number of vibrating atoms is large ($N_T \cong 10^{22}$ cm^{-3}), $l_T \cong 10^{-6}$ cm = 100 A.

The above estimates are purely qualitative, still they give a correct idea of the relation between mean free paths resulting from scattering by centres of different nature.

Indeed, experiment shows that *the main part in scattering at high temperature is played by the thermal vibrations of the lattice. As the temperature is decreased the mobility begins to be determined by impurity ion scattering*. If the impurity ion concentration is small the main part will be played by impurity atom or defect scattering. Below we will be able to arrive at these conclusions with the aid of mathematical expressions for relaxation times produced by different types of scattering centres.

54. RELATIONSHIP BETWEEN RELAXATION TIME AND EFFECTIVE CROSS SECTION

All kinetic phenomena bear some relationship to relaxation times averaged over energy values. The relaxation time is determined by the collision integral the value of which, in turn, depends on the probabilities of electron and hole transitions from state to state as a result of collisions. Therefore, it would be natural to suppose that there must be a connection between the effective scattering cross section and the relaxation time. A proof is presented in quantum mechanics and in physical kinetics of the proposition that *elementary probabilities of direct and reverse transitions are equal*. This means that $w(\mathbf{k}, \mathbf{k}') = w(\mathbf{k}', \mathbf{k})$. With this equality in mind write the expression for the relaxation time:

$$\left(\frac{\partial f}{\partial t}\right)_c = -\frac{f-f_0}{\tau(\mathbf{k})} = \frac{1}{4\pi^3} \int\limits_{(V_{\mathbf{k}})} \{w(\mathbf{k}', \mathbf{k}) f(\mathbf{k}')[1-f(\mathbf{k})] - w(\mathbf{k}, \mathbf{k}') f(\mathbf{k})[1-f(\mathbf{k}')]\} \, d\tau \mathbf{k}'. \qquad (54.1)$$

Taking account of the fact that the collision integral turns zero for the equilibrium function $f = f_0$, setting $f = f_0 + f^{(1)}$ and neglecting terms of the $f^{(1)^2}$ type we obtain

$$\frac{1}{\tau(\mathbf{k})}=-\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})}\left\{w(\mathbf{k}',\mathbf{k})\left[\frac{f^{(1)}(\mathbf{k}')}{f^{(1)}(\mathbf{k})}[1-f_0(\mathbf{k})]-f_0(\mathbf{k}')\right]-\right.$$
$$\left.-w(\mathbf{k},\mathbf{k}')\left[1-f_0(\mathbf{k}')-\frac{f^{(1)}(\mathbf{k}')}{f^{(1)}(\mathbf{k})}f_0(\mathbf{k})\right]\right\}d\tau_{\mathbf{k}'}, \qquad (54.2)$$

which coincides with the expression (36.6s).

Making use of the equality $w(\mathbf{k},\mathbf{k}')=w(\mathbf{k}',\mathbf{k})$ and cancelling out similar terms in (54.2) we represent the expression for $\tau(\mathbf{k})$ in the form

$$\frac{1}{\tau(\mathbf{k})}=\frac{1}{4\pi^3}\int w(\mathbf{k},\mathbf{k}')\left[1-\frac{f^{(1)}(\mathbf{k}')}{f^{(1)}(\mathbf{k})}\right]d\tau_{\mathbf{k}'}. \qquad (54.3)$$

According to (37.16) $f^{(1)}$ should be of the form

$$f^{(1)}(\mathbf{k})=-\left.\frac{\partial f_0}{\partial E}\right|_{\mathbf{k}}(\mathbf{X}(\mathbf{k})\,\mathbf{v}(\mathbf{k}));\quad f^{(1)}(\mathbf{k}')=-\left.\frac{\partial f_0}{\partial E}\right|_{\mathbf{k}'}(\mathbf{X}(\mathbf{k}')\,\mathbf{v}(\mathbf{k}')). \qquad (54.4)$$

Substituting the expression (54.4) for $f^{(1)}(\mathbf{k})$ and $f^{(1)}(\mathbf{k}')$ we obtain the equation for $\tau(\mathbf{k})$:

$$\frac{1}{\tau(\mathbf{k})}=\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})}w(\mathbf{k},\mathbf{k}')\left[1-\frac{\left.\frac{\partial f_0}{\partial E}\right|_{\mathbf{k}'}(\mathbf{X}(\mathbf{k}')\,\mathbf{v}(\mathbf{k}'))}{\left.\frac{\partial f_0}{\partial E}\right|_{\mathbf{k}}(\mathbf{X}(\mathbf{k})\,\mathbf{v}(\mathbf{k}'))}\right]d\tau_{\mathbf{k}'}. \qquad (54.5)$$

The quantity $\mathbf{X}$ is to be determined from the equations (37.52) which show that $\mathbf{X}$, in turn, depends in a complex way on the relaxation time, and for this reason the equation (54.5) turns out to be a nonlinear integral equation which cannot be solved in a general form. The introduction of the relaxation time enabled the solution of the kinetic equation to be obtained; however, the difficulties connected with the solution of the Boltzmann equation have not disappeared with the introduction of the relaxation time but have simply been carried over from the equation for $f^{(1)}$ to that for τ. Moreover, in solving the equation for $f^{(1)}$ we have assumed that the relaxation time is not affected by external fields. To be more precise, we may say that the scattering processes take a different course in the absence or in the presence of the fields $\mathbf{E}$, $\mathbf{B}$. Below we shall see that the dependence of the scattering processes on the external fields makes itself manifest in the form of certain physical phenomena. However, since we assumed that $\tau(\mathbf{k})$ is independent of the external fields, we may formally consider the equation (54.5) for the case of $\mathbf{E}\rightarrow 0$, $\mathbf{B}\rightarrow 0$.

In this case

$$\frac{1}{\tau(\mathbf{k})}=\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})} w(\mathbf{k},\mathbf{k}')\left[1-\frac{\left.\frac{\partial f_0}{\partial E}\right|_{\mathbf{k}'}(\mathbf{v}(\mathbf{k}')\mathbf{E}^0)\,\tau(\mathbf{k}')}{\left.\frac{\partial f_0}{\partial E}\right|_{\mathbf{k}}(\mathbf{v}(\mathbf{k})\mathbf{E}^0)\,\tau(\mathbf{k})}\right]d\tau_{\mathbf{k}'}, \tag{54.6}$$

where $\mathbf{E}^0$ is a unit vector, generally, of arbitrary direction.

For most practically important cases the expression (54.6) may be appreciably simplified. It was assumed previously that τ depends only on the energy. Orient the z-axis along the $\mathbf{k}$ vector; then $\mathbf{k}'$ will determine the direction of the particle motion after scattering. Choosing $\mathbf{E}^0$ along the z-axis we obtain

$$(\mathbf{v}(\mathbf{k})\,\mathbf{E}^0)=v(\mathbf{k});\quad (\mathbf{v}(\mathbf{k}')\,\mathbf{E}^0)=v(\mathbf{k}')\cos(\mathbf{k}',\mathbf{k})=v(\mathbf{k}')\cos\theta \tag{54.7}$$

and, in consequence,

$$\frac{1}{\tau(\mathbf{k})}=\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}'})} w(\mathbf{k},\mathbf{k}')\left[1-\frac{\left.\frac{\partial f_0}{\partial E}\right|_{\mathbf{k}'}\tau(E')\,v(\mathbf{k}')\cos(\mathbf{k}',\mathbf{k})}{\left.\frac{\partial f_0}{\partial E}\right|_{\mathbf{k}}\tau(E)\,v(\mathbf{k})}\right]d\tau_{\mathbf{k}'}. \tag{54.8}$$

Consider the simplest case of spherical constant-energy surfaces. As we have seen in Sec. 53, an elastic collision of a particle with a scattering centre results in a negligible variation of its energy E by the amount not exceeding $\frac{4m}{M}E$. Hence, we may presume that an elastic scattering retains the particle state in the Brillouin zone on the same constant-energy surface. In these assumptions the expression (54.6) may be written in the form

$$\frac{1}{\tau(\mathbf{k})}=\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})} w(\mathbf{k},\mathbf{k}')\,[1-\cos(\mathbf{k}',\mathbf{k})]\,d\tau_{\mathbf{k}'}. \tag{54.9}$$

Since $E\cong E'$, the modulus of velocity remains unchanged:

$$\tau(E')\cong\tau(E)\quad\text{and}\quad v(\mathbf{k}')\cong v(\mathbf{k}).$$

In the spherical coordinate system with a polar axis directed along the vector $\mathbf{k}$, $\cos(\mathbf{k}',\mathbf{k})=\cos\theta$ and

$$\frac{1}{\tau(\mathbf{k})}=\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})} w(\mathbf{k},\mathbf{k}')\,[1-\cos\theta]\,d\tau_{\mathbf{k}'}=$$

$$=\frac{1}{4\pi^3}\int w(k,k',\theta,\varphi)\,(1-\cos\theta)\,k'^2\,dk'\,d\Omega_{\mathbf{k}}. \tag{54.10}$$

Represent $w(k,k',\theta,\varphi)$ in the form of two functions: the radial $\tilde{w}(k,k')$, and the angular $\overset{\approx}{w}(\theta,\varphi)$. Suppose $\overset{\approx}{w}(\theta,\varphi)$ *is the probability* *of the particle being scattered into a unit solid angle in the direction* (θ, φ). The differential effective cross section is the probability for a particle belonging to a single particle flow to be scattered by one scattering centre; therefore,

$$\tilde{\tilde{w}}(\theta, \varphi) = Nv(\mathbf{k})\,\sigma(\theta, \varphi), \qquad (54.11)$$

since the flow $1 \cdot v(\mathbf{k})$ corresponds to one particle with the velocity $v(\mathbf{k})$.

Let $\tilde{w}(k, k')$ be *the probability of transition of a particle from the unit volume* 1 (**k**) *into a spherical layer of the radius* k' *and of unit thickness* $(dk' = 1)$.

Consider the product $\tilde{w}(k, k')\,\tilde{\tilde{w}}(\theta, \varphi)$; it represents *the probability of transition into a volume cut out by a unit solid angle from a spherical layer of unit thickness.* The base area of this volume is, evidently, equal to k'^2. Should we divide $\tilde{w}(k, k')\,\tilde{\tilde{w}}(\theta, \varphi)$ by k'^2 we would obtain the probability of transition into a layer of unit thickness resting on a base of unit area, i.e. into a unit volume with its centre at the point $\mathbf{k}'$. Therefore, we may write

$$w(\mathbf{k}; \mathbf{k}') = \frac{\tilde{w}(k, k')\,\tilde{\tilde{w}}(\theta, \varphi)}{k'^2} = \frac{\tilde{w}(k, k')\,Nv(\mathbf{k})\,\sigma(\theta, \varphi)}{k'^2}. \qquad (54.12)$$

Express the relaxation time in terms of the effective cross section:

$$\frac{1}{\tau(\mathbf{k})} = Nv(\mathbf{k}) \int\limits_{(4\pi)} \sigma(\theta, \varphi)\,(1 - \cos\theta)\,d\Omega \cdot \frac{1}{4\pi^3} \int\limits_0^\infty \tilde{w}(k, k')\,dk'. \qquad (54.13)$$

In case of elastic scattering the state of the scattered particle remains on the constant-energy surface, therefore $\tilde{w}(k, k')$ should be proportional to $\delta(k - k')$. Take $4\pi^3$ for the proportionality factor:

$$\tilde{w}(k, k') = 4\pi^3\delta\,(k - k') \qquad (54.14)$$

$$\frac{1}{4\pi^3} \int\limits_0^\infty 4\pi^3\delta\,(k - k')\,dk' = 1, \qquad (54.15)$$

since the scattering resulting in any final state k' is a certainty.

Taking into account (54.15) we write finally

$$\frac{1}{\tau(\mathbf{k})} = Nv(\mathbf{k}) \int\limits_{(4\pi)} \sigma(\theta, \varphi)\,(1 - \cos\theta)\,d\Omega = Nv(\mathbf{k})\,\sigma_c, \qquad (54.16)$$

or

$$\tau(\mathbf{k}) = \frac{1}{Nv(\mathbf{k})\,\sigma_c}, \qquad l(\mathbf{k}) = \tau(\mathbf{k})\,v(\mathbf{k}) = \frac{1}{N\sigma_c}, \qquad (54.17)$$

which is analogous to (53.8)

Summary of Secs. 53-54

1. The differential scattering cross section $\sigma(\theta, \varphi)$ is the probability for one particle of unit velocity to fall, after a collision with one scattering centre, inside a unit solid angle $d\Omega = 1$ built around the direction θ, φ with the scattering centre at the apex. Knowing $\sigma(\theta, \varphi)$ we may find the number of particles $dn'(\theta, \varphi, x)$ falling inside a solid angle $d\Omega(\theta, \varphi)$ when particles, belonging to a flow with particle concentration n_1 and velocity v, are scattered by centres whose concentration is N over the path dx:

$$-dn'(\theta, \varphi, x) = \sigma(\theta, \varphi)\, n_1(x)\, vN\, dx\, d\Omega(\theta, \varphi). \qquad (54.1\text{s})$$

2. The integral effective scattering cross section S^* is the term applied to the probability of scattering through arbitrary angles. S^* is related to $\sigma(\theta, \varphi)$ by the normalizing condition:

$$S^* = \int\limits_{(4\pi)} \sigma(\theta, \varphi)\, d\Omega. \qquad (54.2\text{s})$$

3. The transport effective cross section, or mobility, or conductivity, effective cross section are the terms applied to the quantity σ_c obtained from the differential cross section $\sigma(\theta, \varphi)$ by integrating over all the angles with a weighting function $(1-\cos\theta)$:

$$\sigma_c = \int\limits_0^{\pi} \int\limits_0^{2\pi} \sigma(\theta, \varphi)(1-\cos\theta)\sin\theta\, d\theta\, d\varphi. \qquad (54.3\text{s})$$

If the differential effective cross section is independent of the angles θ and φ, then

$$S^* = 4\pi\sigma = \sigma_c. \qquad (54.4\text{s})$$

Generally, the ratio S^* to σ_c is equal to the average number of collisions q which result in complete loss of directional motion:

$$q = \frac{v_z}{\langle \Delta v_z \rangle} = \frac{S^*}{\sigma_c}. \qquad (54.5\text{s})$$

In case of isotropic scattering the directional velocity is lost, on the average, after one collision.

4. The mean free path l_i for directional motion is related to σ_{ci} and to the concentration of scattering centres N_i:

$$l_i = \frac{1}{N_i \sigma_{ci}}; \qquad (54.6\text{s})$$

$$l^{-1} = \sum_i N_i \sigma_{ci} = \sum_i l_i^{-1}. \qquad (54.7\text{s})$$

5. The relaxation time $\tau(\mathbf{k})$ is related to σ_c:

$$\frac{1}{\tau(\mathbf{k})} = Nv(\mathbf{k})\,\sigma_c. \tag{54.8s}$$

6. The variation of energy of a particle of mass m* as a result of scattering by a centre of mass M is, on the everage, equal to

$$\langle \Delta E \rangle = 2\,\frac{m^*}{M}\,E\,\frac{\sigma_c}{S^*}\,. \tag{54.9s}$$

7. Electrons and holes can be scattered by: (1) impurity ions; (2) impurity atoms; (3) vacancies and other point defects; (4) dislocations; (5) crystal surfaces, cleavage planes, grain boundaries; (6) electrons and holes; (7) thermal lattice vibrations.

8. The task of the theory of scattering is to calculate the effective scattering cross section for centres of various nature. This will enable $\tau(\mathbf{k})$ and $\langle E^r\tau^s\rangle$ to be calculated from σ_c, and the kinetic phenomena depending on the scattering mechanism to be described.

55. ELEMENTS OF QUANTUM TRANSITION THEORY

As was indicated in the preceding paragraph, to calculate the kinetic coefficients, the differential effective cross section should be known. The latter may be calculated with the aid of quantum mechanical methods.

Since we shall need some results of the quantum transition theory not only for the scattering theory but also for the theory of optical phenomena we would like to recall its fundamentals.

Suppose a quantum system described by a Hamiltonian $\hat{\mathbf{H}}^0$ has a set of stationary states characterized by the energy $E^0(\alpha)$ and the wave functions $\psi^0_\alpha(\mathbf{r})$ satisfying the equation

$$\hat{\mathbf{H}}^0\psi^0_\alpha(\mathbf{r}) = E^0(\alpha)\,\psi^0_\alpha(\mathbf{r}) \tag{55.1}$$

(α is a full set of physical quantities). The wave function $\psi^0_\alpha(\mathbf{r})$ is a solution of the *stationary* Schrödinger equation. The solution of the *time-dependent Schrödinger equation*

$$i\hbar\,\frac{\partial\psi^0_\alpha(\mathbf{r},t)}{\partial t} = \hat{\mathbf{H}}^0\psi^0_\alpha(\mathbf{r},t) \tag{55.2}$$

is connected with the solution of the stationary equation by a simple relation

$$\psi^0_\alpha(\mathbf{r},t) = \psi^0_\alpha(\mathbf{r})\,e^{-i\frac{E^0(\alpha)}{\hbar}t}, \tag{55.3}$$

which is easily checked by directly substituting (55.3) into (55.2) to obtain (55.1).

Suppose at $t=0$ a field $\hat{\mathbf{W}}=\hat{\mathbf{W}}(\mathbf{r}, t)$ has been applied to the quantum system. Now the system Hamiltonian assumes the form

$$\hat{\mathbf{H}}=\hat{\mathbf{H}}^{0}+\hat{\mathbf{W}}. \tag{55.4}$$

The wave function of the perturbed system $\psi(\mathbf{r}, t)$ is determined from the equation

$$i\hbar\frac{\partial\psi}{\partial t}=\hat{\mathbf{H}}\psi=(\hat{\mathbf{H}}^{0}+\hat{\mathbf{W}})\psi. \tag{55.5}$$

To solve this equation expand, following Dirac, $\psi(\mathbf{r}, t)$ in series of the functions $\psi_{\alpha}^{0}(\mathbf{r}, t)$:

$$\psi(\mathbf{r}, t)=\int c_{\alpha}(t)\,\psi_{\alpha}^{0}(\mathbf{r}, t)\,d\alpha. \tag{55.6}$$

We have written the expansion of $\psi(\mathbf{r}, t)$ in $\psi_{\alpha}^{0}(\mathbf{r}, t)$ in the assumption that the spectrum of α is continuous, otherwise a sum over all possible values of α should be used instead of the integral. Substitute (55.6) into (55.5):

$$i\hbar\int\left\{\frac{dc_{\alpha}(t)}{dt}\psi_{\alpha}^{0}(\mathbf{r}, t)+c_{\alpha}(t)\frac{\partial\psi_{\alpha}^{0}(\mathbf{r}, t)}{\partial t}\right\}d\alpha= \\ =\int c_{\alpha}(t)\left\{\hat{\mathbf{H}}^{0}\psi_{\alpha}^{0}(\mathbf{r}, t)+\hat{\mathbf{W}}\psi_{\alpha}^{0}(\mathbf{r}, t)\right\}d\alpha. \tag{55.7}$$

Taking into account (55.2) we obtain

$$i\hbar\int\frac{dc_{\alpha}(t)}{dt}\psi_{\alpha}^{0}(\mathbf{r}, t)\,d\alpha=\int c_{\alpha}(t)\,\hat{\mathbf{W}}\psi_{\alpha}^{0}(\mathbf{r}, t)\,d\alpha. \tag{55.8}$$

Multiply (55.8) by $\psi_{\beta}^{0*}(\mathbf{r}, t)$ and integrate over the entire space. The normalizing condition for the functions $\psi_{\alpha}^{0}(\mathbf{r})$

$$\int\psi_{\alpha}^{0*}(\mathbf{r})\,\psi_{\alpha}^{0}(\mathbf{r})\,d\tau=\delta(\beta-\alpha) \tag{55.9}$$

enables us to write

$$i\hbar\frac{dc_{\beta}}{dt}=\int c_{\alpha}(t)\,e^{\frac{i}{\hbar}[E^{0}(\beta)-E^{0}(\alpha)]t}\,W_{\beta\alpha}\,d\alpha. \tag{55.10}$$

The equation (55.10) is the equation (55.5) in matrix form. This is a precise equation. The general solution of this equation is extremely difficult, and because of this it is usually solved by the perturbation theory methods. Suppose $\hat{\mathbf{W}}$ is small enough to be considered a perturbation. The magnitude of the matrix element is assumed to be the measure of the magnitude of the physical quantities

$$W_{\beta\alpha}=\int\psi_{\beta}^{0*}(\mathbf{r})\,\hat{\mathbf{W}}(\mathbf{r}, t)\,\psi_{\alpha}^{0}(\mathbf{r})\,d\tau=W_{\beta\alpha}(t). \tag{55.11}$$

We will seek the solution $c_\alpha(t)$ in the form of a series

$$c_\alpha(t) = c_\alpha^{(0)}(t) + c_\alpha^{(1)}(t) + c_\alpha^{(2)}(t) + \ldots, \qquad (55.12)$$

where $c_\alpha^{(0)}(t)$ is the zero approximation corresponding to $\hat{\mathbf{W}} = 0$. In the zero approximation, when $\hat{\mathbf{W}} = 0$, the right-hand side of (55.10) turns zero; therefore, in this approximation the $c_\alpha^{(0)}(t)$ are constants, in full agreement with the fact that for $\hat{\mathbf{W}} = 0$ the system is in a stationary state. Consider the state corresponding to the energy $E^0(\gamma)$. For this state all the $c_\alpha^{(0)} = 0$ for $\alpha \neq \gamma$. Since the probability of the system being in any of the α states is 1, for $c_\alpha^{(0)} = 0$ $(\alpha \neq \gamma)$ it must be put $c_\gamma = \infty$, i.e.

$$c_\alpha^{(0)} = \delta(\alpha - \gamma). \qquad (55.13)$$

To obtain an equation for $c_\beta^{(1)}(t)$ one should take the zero approximation for $c_\alpha(t)$ since the equation (55.10) contains $W_{\beta\alpha}$:

$$i\hbar \frac{dc_\beta^{(1)}(t)}{dt} = \int \delta(\alpha - \gamma)\, e^{\frac{i}{\hbar}[E^0(\beta) - E^0(\alpha)]t}\, W_{\beta\alpha}(t)\, d\alpha =$$

$$= e^{\frac{i}{\hbar}[E^0(\beta) - E^0(\gamma)]t}\, W_{\beta\gamma}. \qquad (55.14)$$

The equation (55.14) may be easily integrated:

$$c_\beta^{(1)}(t) = \frac{1}{i\hbar} \int_0^t W_{\beta\gamma}(t)\, e^{i\omega_{\beta\gamma}t}\, dt, \qquad (55.15)$$

where

$$\omega_{\beta\gamma} = \frac{E^0(\beta) - E^0(\gamma)}{\hbar}.$$

The quantity $|c_\beta^{(1)}(t)|^2$ *is the probability of the quantum system being in the state* $\psi_\beta^0(\mathbf{r}, t)$ *at the moment of time* t. At the moment $t = 0$ the system was in the state $\psi_\gamma^0(\mathbf{r}, 0)$. Using the expression (55.15) we obtain for the probability $|c_\beta^{(1)}(t)|^2$

$$|c_\beta^{(1)}(t)|^2 = \frac{1}{\hbar^2} \left| \int_0^t W_{\beta\gamma}(t)\, e^{i\omega_{\beta\gamma}t}\, dt \right|^2. \qquad (55.16)$$

This quantity may be regarded as *the probability of the system transition from the state* $\psi_\gamma^0(\mathbf{r}, t)$ *to the state* $\psi_\beta^0(\mathbf{r}, t)$ *under the action of the perturbation* $\hat{\mathbf{W}}$ *during the time* t.

The transition probability *per unit time* which we will denote by $w(\gamma, \beta)$ may be determined from the relation

$$w(\gamma, \beta) = \frac{d}{dt} |c_\beta^{(1)}(t)|^2 = \frac{d}{dt} \frac{1}{\hbar^2} \left| \int_0^t W_{\beta\gamma} e^{i\omega_{\beta\gamma}t}\, dt \right|^2. \qquad (55.17)$$

Consider a perturbation independent of time. Taking the matrix element out of the integral sign and integrating over time we obtain

$$w(\gamma,\beta)=\frac{|W_{\beta\gamma}|^2}{\hbar^2\omega_{\beta\gamma}^2}\frac{d}{dt}|e^{i\omega_{\beta\gamma}t}-1|^2=\frac{|W_{\beta\gamma}|^2}{\hbar^2\omega_{\beta\gamma}^2}\frac{d}{dt}2(1-\cos\omega_{\beta\gamma}t)=$$
$$=\frac{2|W_{\beta\gamma}|^2}{\hbar^2\omega_{\beta\gamma}^2}\omega_{\beta\gamma}\sin\omega_{\beta\gamma}t=\frac{2}{\hbar^2}|W_{\beta\gamma}|^2\frac{\sin\omega_{\beta\gamma}t}{\omega_{\beta\gamma}}. \qquad (55.18)$$

The expression (55.Г8) shows the probability of transition from the state γ to the state β to be oscillating. The oscillation frequency depends on the difference in energies. If $w(\gamma,\beta)$ is considered to be a function of $\omega_{\beta\gamma}$ it is seen to decrease monotonically with the increase in the frequency for each moment of time. In other words, the transition probability decreases with the increase in the difference of energies between the initial and the final states.

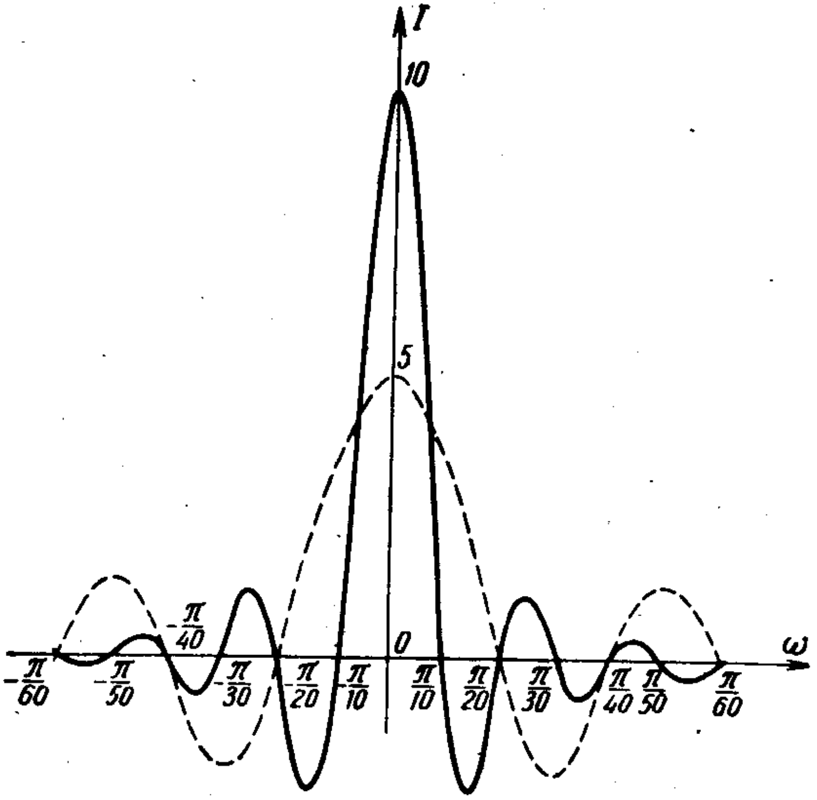


Fig. 79. The frequency dependence of $I=\frac{\sin\omega t}{\omega}$

Consider the quantity $I=\frac{\sin\omega_{\beta\gamma}t}{\omega_{\beta\gamma}}=t\frac{\sin x}{x}$ as a function of the frequency. If $\omega_{\beta\gamma}\to 0$ then

$$\lim_{\omega_{\beta\gamma}\to 0} I=\lim_{\omega_{\beta\gamma}\to 0}\frac{\sin\omega_{\beta\gamma}t}{\omega_{\beta\gamma}}=t. \qquad (55.19)$$

Figure 79 shows the graph of the function I.

The greater the time t, the greater the maximum and the less the frequency interval spanned by the first maximum:

$$\Delta\omega_{\beta\gamma} = \frac{2\pi}{t}. \tag{55.20}$$

For $t \longrightarrow \infty$ the interval of the first maximum $\Delta\omega_{\beta\gamma} \longrightarrow 0$ and $I \longrightarrow \infty$.

Consider the integral of I over the frequencies:

$$\int_{-\infty}^{\infty} I d\,\omega_{\beta\gamma} = \int_{-\infty}^{\infty} \frac{\sin \omega_{\beta\gamma} t}{\omega_{\beta\gamma}}\, d\omega_{\beta\gamma} = \pi. \tag{55.21}$$

Thus, for large interval of t

$$\frac{1}{\pi} \frac{\sin \omega_{\beta\gamma} t}{\omega_{\beta\gamma}} \cong \delta(\omega_{\beta\gamma}). \tag{55.22}$$

The equality (55.22) *will be valid if the perturbation acts for a sufficiently long time.* In this case we may write

$$w(\gamma, \beta) = \frac{2\pi}{\hbar^2} |W_{\beta\gamma}|^2 \delta(\omega_{\beta\gamma}) = \frac{2\pi}{\hbar} |W_{\beta\gamma}|^2 \delta[E^0(\beta) - E^0(\gamma)]. \tag{55.23}$$

The expression (55.23) shows that *the transition probability is proportional to the square of the modulus of the corresponding element of the perturbation matrix. It is non-zero if $E^0(\beta) = E^0(\gamma)$, i.e. if the energy of the system is conserved.*

Use the expression (55.23) to calculate the probability of transition from one state to another when particles are scattered by a force centre, the potential energy of interaction of particles with which is described by the function $V(\mathbf{r}) = \hat{\mathbf{W}}(\mathbf{r})$. To calculate the matrix element of the perturbation operator one should take the Bloch functions for the initial and the final state to obtain

$$W_{\mathbf{k}'\mathbf{k}} = \int \psi_{\mathbf{k}'}^{0*}(\mathbf{r})\, \hat{\mathbf{W}}(\mathbf{r})\, \psi_{\mathbf{k}}^{0}(\mathbf{r})\, d\tau = \int e^{-i(\mathbf{k}'-\mathbf{k},\, \mathbf{r})} V(\mathbf{r})\, \varphi_{\mathbf{k}'}^{*}(\mathbf{r})\, \varphi_{\mathbf{k}}(\mathbf{r})\, d\tau. \tag{55.24}$$

In the effective mass approximation the Bloch function may be replaced by the de Broglie wave, i.e. the periodic functions $\varphi_{\mathbf{k}'}(\mathbf{r})$ and $\varphi_{\mathbf{k}}(\mathbf{r})$ are substituted by constants. Choose the amplitudes of the incident and scattered waves from the following considerations. *The scattered wave* $\psi_{\mathbf{k}'}(\mathbf{r})$ describes the particle position in the entire space and, for this reason, should be normalized to a δ-function to be written in the form

$$\psi_{\mathbf{k}'}(\mathbf{r}) = \frac{1}{(2\pi)^{3/2}} e^{i(\mathbf{k}'\mathbf{r})}. \tag{55.25}$$

Normalize the *incident wave* so that it would describe a unit flow of probability. Since the density of the probability flow is

$$\mathbf{j}_w = \frac{i\hbar}{2m^*}(\psi\nabla\psi^* - \psi^*\nabla\psi), \tag{55.26}$$

putting $\psi_{\mathbf{k}}(\mathbf{r}) = Ae^{i(\mathbf{kr})}$ we obtain for $\mathbf{j}_w$

$$\mathbf{j}_w = \frac{i\hbar}{2m^*}[\mathrm{A}^*(-i\mathbf{k})\,\mathrm{A} - \mathrm{A}^*(i\mathbf{k})\,\mathrm{A}] = \frac{|\mathrm{A}|^2}{m^*}\hbar\mathbf{k} = |\mathrm{A}|^2\,\mathbf{v}. \tag{55.27}$$

Thus, *the flow of probability of unit density satisfies the condition*

$$|\mathrm{A}|^2 v = 1; \quad \mathrm{A} = \sqrt{\frac{1}{v}} = \sqrt{\frac{m^*}{\hbar k}} \tag{55.28}$$

or

$$\psi_{\mathbf{k}}(\mathbf{r}) = \left(\frac{m^*}{\hbar k}\right)^{1/2} e^{i(\mathbf{kr})}. \tag{55.29}$$

Substituting $\psi_{\mathbf{k}}(\mathbf{r})$ and $\psi_{\mathbf{k}'}(\mathbf{r})$ from (55.29) and (55.25) into (55.24) we obtain for the matrix element of the perturbation operator

$$W_{\mathbf{k}'\mathbf{k}} = \int \frac{1}{(2\pi)^{3/2}} e^{-i(\mathbf{k}'\mathbf{r})} V(\mathbf{r}) \left(\frac{m^*}{\hbar k}\right)^{1/2} e^{i(\mathbf{kr})}\, d\tau =$$

$$= \frac{m^{*1/2}}{(2\pi)^{3/2}(\hbar k)^{1/2}} \int e^{i(\varkappa\mathbf{r})} V(\mathbf{r})\, d\tau, \tag{55.30}$$

where

$$\varkappa = \mathbf{k} - \mathbf{k}'.$$

Denote the integral contained in the matrix element (55.30) by $V_\varkappa$

$$V_\varkappa = \int e^{i(\varkappa\mathbf{r})} V(\mathbf{r})\, d\tau. \tag{55.31}$$

Consider the case of a spherically symmetrical perturbation: $V(\mathbf{r}) = V(r)$. To calculate (55.31) it will be convenient to adopt a spherical coordinate system with the polar axis directed along the $\varkappa$ vector:

$$V_\varkappa = \int_0^\infty \int_0^\pi \int_0^{2\pi} e^{i\varkappa r\cos\theta'} V(r)\, r^2\, dr \sin\theta'\, d\theta'\, d\varphi =$$

$$= 2\pi \int_0^\infty r^2 V(r)\, dr \int_0^\pi e^{i\varkappa r\cos\theta'} \sin\theta'\, d\theta'. \tag{55.32}$$

The integral over θ' is easily calculated:

$$\int_0^\pi e^{i\varkappa r\cos\theta'} \sin\theta'\, d\theta' = -\frac{1}{i\varkappa r} e^{i\varkappa r\cos\theta'} \Big|_0^\pi = \frac{2\sin\varkappa r}{\varkappa r}. \tag{55.33}$$

Substituting (55.33) into (55.30) we obtain

$$V_\varkappa = 4\pi \int_0^\infty rV(r)\,\frac{\sin \varkappa r}{\varkappa}\,dr. \qquad (55.34)$$

Write the expression (55.23) for the probability of transition using (55.34):

$$w(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar}\,\frac{m^*}{(2\pi)^3\hbar k}\,|V_\varkappa|^2\,\delta[E(\mathbf{k})-E(\mathbf{k}')], \qquad (55.35)$$

Should (55.35) be integrated over all values of $\mathbf{k}'$ we should obtain the probability of transition $w(\mathbf{k})$ of one particle constituting a flow of unit density from a given state $\mathbf{k}$ to all the possible states. Therefore, $w(\mathbf{k})$ coincides with the integral effective cross section:

$$w(\mathbf{k}) = S^* = \int w(\mathbf{k},\mathbf{k}')\,d\tau_{\mathbf{k}'} = \frac{m^*}{4\pi^2\hbar^3 k}\int |V_\varkappa|^2\,\delta[E(\mathbf{k})-E(\mathbf{k}')]\,d\tau_{\mathbf{k}'}. \qquad (55.36)$$

The integral over $\mathbf{k}'$ may be easily calculated if it is taken into account that

$$\delta\left[\frac{\hbar^2k^2}{2m^*}-\frac{\hbar^2k'^2}{2m^*}\right] = \frac{2m^*}{\hbar^2}\,\delta(k^2-k'^2);$$

$$d\tau_{\mathbf{k}'} = d\Omega\,\frac{k'}{2}\,d(k'^2). \qquad (55.37)$$

The integration over $\mathbf{k}'$ reduces to the substitution of k for k':

$$S^* = \frac{m^{*2}k}{4\pi^2\hbar^4 k}\int_{(4\pi)} |V_\varkappa|^2_{k=k'}\,d\Omega. \qquad (55.38)$$

On the other hand,

$$S^* = \int \sigma(\theta,\varphi)\,d\Omega. \qquad (55.39)$$

Comparing (55.39) with (55.38) we write

$$\sigma(\theta,\varphi) = \frac{m^{*2}}{4\pi^2\hbar^4}\left|4\pi\int_0^\infty rV(r)\,\frac{\sin\varkappa r}{\varkappa}\,dr\right|^2 = \left|\frac{2m^*}{\hbar^2}\int_0^\infty rV(r)\,\frac{\sin\varkappa r}{\varkappa}\,dr\right|^2. \qquad (55.40)$$

Hence, having obtained $V_\varkappa$ we may express *the differential effective cross section with the aid of the perturbation matrix element.*

Summary of Sec. 55

1. The perturbation applied to a quantum system formerly in a stationary state results in transitions from this state into other stationary states. The probability of transition from the state α

to the state β per unit time is

$$w(\alpha, \beta) = \frac{d}{dt}\frac{1}{\hbar^2}\left|\int_0^t W_{\beta\alpha}(t)\, e^{-i\frac{E(\beta)-E(\alpha)}{\hbar}t}\, dt\right|^2. \qquad (55.1s)$$

If the perturbation is independent of time, then

$$w(\alpha, \beta) = \frac{2\pi}{\hbar}\,|W_{\beta\alpha}|^2\,\delta\left[E(\beta) - E(\alpha)\right]. \qquad (55.2s)$$

The matrix element $W_{\beta\alpha}$ is calculated with the aid of the wave functions for stationary states $\psi_\beta^0(\mathbf{r})$ and $\psi_\alpha^0(\mathbf{r})$.

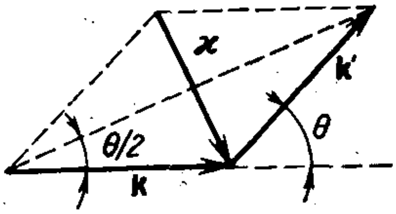


Fig. 80. The relation between ϰ and the scattering angle θ

2. For a spherically symmetrical perturbation the transition probability (55.2s) enables the differential effective scattering cross section to be expressed in the form

$$\sigma(\theta, \varphi) = \left|\frac{2m^*}{\hbar^2}\int_0^\infty rV(r)\,\frac{\sin\varkappa r}{\varkappa}\,dr\right|^2, \qquad (55.3s)$$

where $V(r)$ describes the electron energy in the field of a spherically symmetrical scattering centre, and $\varkappa = |\mathbf{k} - \mathbf{k}'|$. Since, according to (55.37), $k = k'$ it follows

$$\varkappa = 2k\sin\frac{\theta}{2}, \qquad (55.4s)$$

where θ is the scattering angle (Fig. 80)

56. IMPURITY ION SCATTERING

Apply the general theory of quantum transitions of the preceding paragraph to determine the effective scattering cross section of electrons and holes by impurity ions. To this end the perturbation energy should be written in the form

$$V(r) = \pm\frac{Ze^2}{\varepsilon r}. \qquad (56.1)$$

The plus sign holds for the case of similar charges of the scattered particle and the ion, the minus sign—for the case of opposite charges. We denoted the ion charge by Ze and the distance between the ion and the particle by r. To generalize the results we will consider the scattering of particles by the centres with a *screening potential*

$$V(r)=\pm\frac{Ze^2}{\varepsilon r}e^{-k_0 r}. \qquad (56.2)$$

The dimensions of the sphere of action of the scattering centre are determined by the quantity $R=\frac{1}{k_0}$. For $k_0=0$, or $R=\infty$, the corresponding field will be the Coulomb field of the charge Ze. The expression for the potential energy (56.2) satisfactorily describes the behaviour of an electron (or a hole) in the field of a neutral atom, k_0 being in this case approximately equal to 10^{8} cm^{-1}

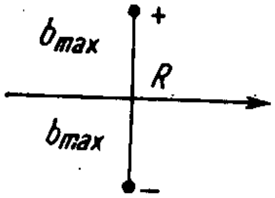


Fig. 81. Maximum impact parameter for a particle moving in the field of two ions

To find the differential effective scattering cross section one should know $V_\varkappa$:

$$\frac{V_\varkappa}{4\pi}=\pm\int\limits_0^\infty \frac{Ze^2}{\varepsilon r}e^{-k_0 r}\frac{r\sin\varkappa r}{\varkappa}\,dr=$$

$$=\pm\frac{Ze^2}{2i\varepsilon\varkappa}\left\{\int\limits_0^\infty\left[e^{-k_0 r+i\varkappa r}-e^{-k_0 r-i\varkappa r}\right]dr\right\}=\pm\frac{Ze^2}{\varepsilon}\,\frac{1}{k_0^2+\varkappa^2}. \qquad (56.3)$$

Substituting the expression (56.3) into (55.40) write for $\sigma(\theta, \varphi)$

$$\sigma(\theta, \varphi)=\frac{4m^{*2}}{\hbar^4}\,\frac{Z^2e^4}{\varepsilon^2}\,\frac{1}{[k_0^2+\varkappa^2]^2}. \qquad (56.4)$$

The angular dependence $\sigma(\theta, \varphi)$ is contained in $\varkappa$:

$$\sigma(\theta, \varphi)=4\left(\frac{Ze^2m^*}{\varepsilon\hbar^2}\right)^2\frac{1}{\left[k_0^2+4k^2\sin^2\frac{\theta}{2}\right]^2}=\sigma(\theta). \qquad (56.5)$$

For the case of ion scattering the expression (56.5) is simplified: $k_0=0$ and

$$\sigma(\theta)=\frac{1}{4}\left(\frac{Ze^2}{\varepsilon m^* v^2}\right)^2\frac{1}{\sin^4\frac{\theta}{2}}. \qquad (56.6)$$

This is the well-known *Rutherford formula* he used to describe the scattering of α-particles by nuclear fields.

Should we calculate S^* and σ_c from (56.6), infinity would be the result. The cause of this is that large effective cross sections correspond to small deflection angles since small deflection angles correspond to large distances between the ion and the particle being scattered, to the so-called *impact parameter b*. For a particle moving in a solid it is not necessary to consider variations of b from zero to infinity. The following simple consideration will enable the upper limit of the impact parameter to be determined. Suppose we have two ions at a distance R from each other (Fig. 81). Obviously, the deflection of a charge carrier is caused by the ion nearest to it. It would be natural to choose $R/2$ as the upper limit of the impact parameter. For a concentration of ions N_I the mean separation between them will be $N_I^{-1/3}$; therefore, the upper limit b_{max} is equal to $\frac{1}{2}N_I^{-1/3}$.

The corresponding minimum deflection angle θ_{min} is determined from the condition

$$\tan\frac{\theta_{min}}{2}=\frac{Ze^2}{\varepsilon m^* v^2}\,\frac{2}{N_I^{-1/3}}. \tag{56.7}$$

The introduction of θ_{min} enables a finite value for S^* to be obtained:

$$S^*=2\pi\int\limits_{\theta_{min}}^{\pi}\sigma(\theta)\sin\theta\,d\theta=\frac{\pi}{2}\,\frac{1}{4}\left(\frac{Ze^2}{\varepsilon m^* v^2}\right)^2\int\limits_{\theta_{min}}^{\pi}\frac{\sin\theta\,d\theta}{\sin^4\frac{0}{2}}=$$

$$=\frac{\pi}{2}\left(\frac{Ze^2}{\varepsilon m^* v^2}\right)^2\int\limits_{\theta_{min}}^{\pi}\frac{2\cos\frac{\theta}{2}\cdot 2d\frac{\theta}{2}}{\sin^3\frac{\theta}{2}}=\pi\left(\frac{Ze^2}{\varepsilon m^* v^2}\right)^2\left(\frac{1}{\sin^2\frac{\theta_{min}}{2}}-1\right). \tag{56.8}$$

Find the effective cross section of conductivity σ_c:

$$\sigma_c=2\pi\int\limits_{\theta_{min}}^{\pi}\sigma(\theta)(1-\cos\theta)\sin\theta\,d\theta=$$

$$=\frac{\pi}{2}\left(\frac{Ze^2}{\varepsilon m^* b^2}\right)^2\int\limits_{\theta_{min}}^{\pi}\frac{(1-\cos\theta)\sin\theta\,d\theta}{\sin^4\frac{\theta}{2}}. \tag{56.9}$$

Taking into account that

$$1-\cos\theta=2\sin^2\frac{\theta}{2};\quad \sin\theta=2\sin\frac{\theta}{2}\cos\frac{\theta}{2}, \tag{56.10}$$

we obtain

$$\int_{\theta_{min}}^{\pi} \frac{(1-\cos\theta)\sin\theta\, d\theta}{\sin^4\frac{\theta}{2}} = 8\int_{\theta_{min}}^{\pi} \frac{\cos\frac{\theta}{2}\cdot d\frac{\theta}{2}}{\sin\frac{\theta}{2}} = -8\ln\sin\frac{\theta_{min}}{2} \qquad (56.11)$$

and

$$\sigma_c = -4\pi\left(\frac{Ze^2}{\varepsilon m^* v^2}\right)^2 \ln\sin\frac{\theta_{min}}{2}. \qquad (56.12)$$

Express θ_{min} in terms of b_{max} in accordance with (56.7):

$$\frac{1}{\sin^2\frac{\theta_{min}}{2}} = 1+\cot^2\frac{\theta_{min}}{2} = 1+\frac{1}{4N_I^{2/3}}\left(\frac{\varepsilon m^* v^2}{Ze^2}\right)^2. \qquad (56.13)$$

Making use of (56.13), (56.8) and (56.9) we obtain in this case for S^* and σ_c

$$S^* = \pi\left(\frac{N_I^{-1/3}}{2}\right)^2 = \pi b_{max}^2 \qquad (56.14)$$

and

$$\sigma_c = -4\pi\left(\frac{Ze^2}{\varepsilon m^* v^2}\right)^2 \ln\left[1+\frac{N_I^{-2/3}}{4}\left(\frac{\varepsilon m^* v^2}{Ze^2}\right)^2\right]^{-1/2} =$$
$$= 2\pi\left(\frac{Ze^2}{\varepsilon m^* v^2}\right)^2 \ln\left[1+\frac{1}{4N_I^{2/3}}\left(\frac{\varepsilon m^* v^2}{Ze^2}\right)^2\right]. \qquad (56.15)$$

The meaning of the expression for S^* is quite obvious: free carriers interact with an ion at any distance, therefore the integral effective cross section for ion scattering is infinitely large. But if we take into account, that for a large number of ions we may confine ourselves to the impact parameter equal to $b_{max} = \frac{N_I^{-1/3}}{2}$, we shall obtain for the effective cross section a value equal to the area of a circle of the radius b_{max}. Note, that the higher is the impurity ion concentration the more their fields are mutually compensated, and the smaller is their integral carrier scattering cross section. Let $N_I = 10^{16}$ cm^{-3}; in this case $b_{max} \cong 10^{-5}$ cm, and $S^* \cong 10^{-10}$ cm^2, which is much greater than the value of S_I assumed in Sec. 53 as an estimate of the part played by ion scattering. The mean free path corresponding to $N_I = 10^{16}$ cm^{-3} would turn out to be $l_I = 10^{-6}$ cm. This means that the impurity ions play a much greater part in scattering and in determining the mean free path than was assumed by us in Sec. 53.

The dependence of l_I on ion concentration is comparatively weak:

$$l_I = \frac{1}{N_I S_I^*} = \frac{1}{N_I \frac{\pi}{4} N_I^{-2/3}} = \frac{4}{\pi N_I^{1/3}} = \frac{8}{\pi} b_{max}. \tag{56.16}$$

Table 18 shows the values of l_I calculated with the aid of formula (56.16) for some values of N_I.

Table 18

N_I, cm^{-3}	10^{21}	10^{18}	10^{15}	10^{12}
l_I, cm	1.3×10^{-7}	1.3×10^{-6}	1.3×10^{-5}	1.3×10^{-4}

The integral effective cross section is independent of the energy and the velocity of charge carriers, but the differential effective cross section is strongly dependent on the carrier velocity:

$$\sigma(\theta) \sim \frac{1}{v^4}. \tag{56.17}$$

Moreover, as the expression for $\sigma(\theta)$ shows, ion scattering is highly anisotropic: $\sigma(\theta) \sim \frac{1}{\sin^4 \frac{\theta}{2}}$. The dependence of the effective cross section of conductivity σ_c on the velocity is of two forms: the power and the logarithmic. Introduce the notation

$$x = \frac{N_I^{-2/3}}{4}\left(\frac{\varepsilon m^* v^2}{Ze^2}\right)^2; \quad \sigma_c = \frac{\pi}{2} N_I^{-2/3} \frac{\ln(1+x)}{x}. \tag{56.18}$$

For $x \gg 1$

$$\sigma_c \sim \frac{1}{v^4}, \tag{56.19}$$

since the logarithmic dependence is weaker than the power dependence. For $x \ll 1$ the expression (56.15) for σ_c may be simplified:

$$\sigma_c \cong \frac{\pi}{2} N_I^{-2/3} \frac{\ln(1+x)}{x} \cong \frac{2\pi N_I^{-2/3}}{4}; \tag{56.20}$$

i.e. *the effective cross section of conductivity is independent of the velocity and of the energy of the particles being scattered.* Consider the conditions necessary for the inequalities (56.19) or (56.20) to be satisfied; to this end introduce the energy into the expressions for x:

$$x = \left(\frac{\varepsilon m^* v^2}{Ze^2 2}\right)^2 \frac{1}{N_I^{2/3}} = \left(\frac{E}{\frac{Ze^2}{\varepsilon N_I^{-1/3}}}\right)^2. \tag{56.21}$$

But $\frac{Ze^2}{\varepsilon N_I^{-1/3}} = V(2b_{max})$ is the potential energy at a distance $2b_{max}$ from the scattering centre. For $E \ll V\ (2b_{max})$, i.e. for the "slow" carriers, σ_c is independent of their velocity and energy. For $E \gg V\,(2b_{max})$ $\sigma_c \sim \frac{1}{v^4}$ or $\sigma_c \sim \frac{1}{E^2}$; for $E \cong V\,(2b_{max})$ the dependence on v remains, but it is much weaker than in case of high energies. The values of $V\ (2b_{max})$ for various N_I and for $\varepsilon = 10$ and $Z = 1$ are shown in Table 19.

Table 19 shows, besides, the values of temperature T K calculated from the condition $V(2b_{max}) = kT$. The energy interval (0-E)

Table 19

N_I, cm^{-3}	10^{21}	10^{18}	10^{15}	10^{12}	10^{9}
$2b_{max}$, cm	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}
$V\ (2b_{max})$, eV	0.143	0.0143	0.0014	0.0001	0.00001
$T_{max} = \frac{V}{k}$, K	1670	167	16.7	1.7	0.17

and the corresponding temperature range, within which σ_c is independent of the velocity, is the greater the greater is the impurity ion concentration. For $N_I \cong 10^{16}$ and less the interval of energy values where σ_c is independent of E is very small, therefore it may be assumed that $\sigma_c \sim E^{-2}$. Consider now the energy dependence of the relaxation time. According to (54.17) and (56.15)

$$\tau_I(\mathbf{k}) = \frac{1}{N_I \sigma_c v(\mathbf{k})} = \frac{\varepsilon^2 m^{*2} v^3}{2\pi N_I Z^2 e^4 \ln\left[1 + \left(\frac{\varepsilon m^* v^2}{2N_I^{1/3} Ze^2}\right)^2\right]}. \qquad (56.22)$$

This expression for τ_I is termed *Conwell-Weisskopf formula*. It follows from (56.22) that $\tau_I \sim v^3$. Expressing the velocity in terms of energy, $v = \left(\frac{2E}{m^*}\right)^{1/2}$, we obtain

$$\tau_I = \frac{\sqrt{2}\,\varepsilon^2 m^{*1/2} E^{3/2}}{N_I \pi Z^2 e^4 \ln\left[1 + \left(\frac{\varepsilon E}{N_I^{1/3} Ze^2}\right)^2\right]} \cong \tau_0 E^{3/2}. \qquad (56.23)$$

This shows that *when charge carriers are scattered by impurity ions the energy dependence of the relaxation time is of the form* $\tau(E) = \tau_0 E^p$ with $p = 3/2$. The relaxation process takes the longer the higher is the carrier energy. The applicability of this formula was already discussed above. For $E \rightarrow 0$ the expression (56.22) may not be applied and should be replaced by the expression

$$\tau_I = \frac{1}{N_I \sigma_c v} = \frac{1}{N_I \frac{\pi}{2} N_I^{-2/3} v} = \frac{2}{\pi N_I^{1/3} v} = \frac{\sqrt{m^*}}{\sqrt{2}\pi N_I^{1/3}} E^{-1/2}, \quad (56.24)$$

which follows directly from (56.23).

Should the mean free path of the charge carriers with the energy E be defined by the relation $l_I(E) = v\tau_I(E)$ it would follow that

$$l_I(E) \cong \frac{1}{N_I \sigma_c}; \quad l_I \sim E^2; \quad l_I(E) = \text{const} \quad (56.25)$$

foı high and low electron and hole energies, respectively.

Return now to the expression (56.5) and calculate with its aid the transport effective cross section:

$$\sigma_c = 2\pi \int_0^\pi 4\left(\frac{Ze^2 m^*}{\varepsilon\hbar^2}\right)^2 \frac{(1-\cos\theta)\sin\theta\, d\theta}{\left[k_0^2 + 4k^2 \sin^2\frac{\theta}{2}\right]^2} =$$

$$= 32\pi\left(\frac{Ze^2 m^*}{\varepsilon\hbar^2}\right)^2 \int_0^\pi \frac{\sin^2\frac{\theta}{2}\, d\sin^2\frac{\theta}{2}}{\left[k_0^2 + 4k^2 \sin^2\frac{\theta}{2}\right]^2}. \quad (56.26)$$

Taking $4k^2$ out of the integral and introducing a new variable $y = \sin^2\frac{\theta}{2}$ we obtain

$$\sigma_c = 2\pi\left(\frac{Ze^2 m^*}{\varepsilon\hbar^2 k^2}\right)^2 \int_0^1 \frac{y\, dy}{\left[\frac{k_0^2}{4k^2} + y\right]^2}. \quad (56.27)$$

The integral in (56.27) may, for example, be calculated by parts, and we obtain as a result

$$\sigma_c = 2\pi\left(\frac{Ze^2 m^*}{\varepsilon\hbar^2 k^2}\right)^2 \left[\ln\left(1 + \frac{4k^2}{k_0^2}\right) - \frac{4k^2}{k_0^2 + 4k^2}\right]. \quad (56.28)$$

For $k_0 \rightarrow 0$ the expression for $\sigma_c \rightarrow \infty$. The parameter k_0 may be connected with the free carrier concentration if the screening action of the "swarm" of electrons or holes attracted by the ion is taken into account. However, since the result is not different in principle from the expression for σ_c and τ_I given by the Conwell-Weisskopf formula, we shall not analyse the expression (56.28).

Summary of Sec. 56

1. The differential effective cross section for charge carrier scattering in the Coulomb field of an ion is given by the Rutherford formula

$$\sigma(\theta, \varphi) = \frac{1}{4}\left(\frac{Ze^2}{\varepsilon m^* v^2}\right)^2 \frac{1}{\sin^4 \frac{\theta}{2}}. \qquad (56.1s)$$

The scattering is highly anisotropic. Small angle scattering with corresponding long impact parameters prevails. The expressions for S^* and σ_c diverge.

2. Finite values of S^* and σ_c may be obtained if the impact parameters are limited to some value b_{max} or if deflection angles smaller than some preset value θ_{min} are disregarded. The physical basis for the introduction of b_{max} is either the screening action of a "swarm" of free carriers, which turns the Coulomb potential into a screening potential of the form (56.2), or the compensating action of the neighbouring ions. In the first case the expressions for σ and σ_c assume the form

$$\sigma(\theta) = 4\left(\frac{Ze^2 m^*}{\varepsilon \hbar^2}\right)^2 \frac{1}{\left[k_0^2 + 4k^2 \sin^2 \frac{\theta}{2}\right]^2}; \qquad (56.2s)$$

$$\sigma_c(\theta) = 2\pi\left(\frac{Ze^2 m^*}{\varepsilon \hbar^2 k^2}\right)^2 \left[\ln\left(1 + \frac{4k^2}{k_0^2}\right) - \frac{4k^2}{k_0^2 + 4k^2}\right], \qquad (56.3s)$$

where k_0 is related to the concentration of free carriers. In the second case we obtain for σ_c

$$\sigma_c = 2\pi\left(\frac{Ze^2}{\varepsilon m^* v^2}\right)^2 \ln\left[1 + \left(\frac{\varepsilon m^* v^2}{2N_I^{2/3} Ze^2}\right)^2\right]. \qquad (56.4s)$$

3. For sufficiently large carrier energies the effective cross section of conductivity σ_c is inversely proportional to the square of the energy. At the same time for small energies σ_c is independent of energy.

4. The expression for relaxation time in case of impurity ion scattering is of the form

$$\tau_I = \frac{(2m^*)^{1/2} \varepsilon^2}{N_I \pi Z^2 e^4} \frac{E^{3/2}}{\ln\left[1 + \left(\frac{\varepsilon E}{N_I^{1/3} Ze^2}\right)^2\right]}. \qquad (56.5s)$$

It is proportional to the 1/2 power of the effective mass. For large energies $\tau_I \sim E^{3/2}$, for small energies $\tau_I \sim E^{-1/2}$. The energy interval in which $\tau_I \sim E^{-1/2}$ plays a prominent part in the zero temperature range.

57. SCATTERING BY NEUTRAL IMPURITY ATOMS

Charge carrier scattering by neutral impurity atoms may be described with the aid of the expressions (56.5) and (56.28). Since for an impurity atom $k_0 \cong 10^8$ cm^{-1} which corresponds to an energy of $\frac{m}{m^*} \cdot 8$ eV, practically, for all the carriers the quantity k^2 may be neglected as compared to k_0^2. This results in the expressions for $\sigma(\theta)$ and σ_c being simplified:

$$\sigma(\theta) = 4\left(\frac{Ze^2m^*}{\varepsilon\hbar^2k_0^2}\right)^2 \tag{57.1}$$

and

$$\sigma_c = 8\pi\left(\frac{Ze^2m^*}{\varepsilon\hbar^2k_0^2}\right)^2, \tag{57.2}$$

i.e. *the scattering is isotropic, and σ and σ_c are independent of the energy of the particles being scattered.*

Assess the magnitude of σ. Put $m^* = m$; $\varepsilon = 10$; $Z = 1$; $k_0 = 10^8$ cm^{-1} and obtain $\sigma \cong 2 \times 10^{-17}$ cm^2.

The relaxation time for scattering by neutral impurity atoms, according to (56.24), is equal to

$$\tau_n = \frac{1}{N_n v \sigma_c} = \frac{1}{N_n \sqrt{2}\pi}\left(\frac{\varepsilon\hbar^2k_0^2}{Ze^2}\right)^2 \frac{E^{-1/2}}{m^{*3/2}}. \tag{57.3}$$

The expressions (57.2) and (57.3) are valid insofar as the neutral impurity atom or the point defect may be described with the aid of the screening potential (56.2). Note one peculiarity of the σ_c dependence on the dielectric permeability of the substance. For the estimates of σ_c and k_0 we used the value which applies to an isolated atom. If a hydrogen-like atom plays the part of an impurity atom, the dimensions of its electron cloud, as was demonstrated in Sec. 21, become ε times greater; therefore, $k_0' = \frac{k_0}{\varepsilon}$ should be substituted for k_0 with the result that the form of σ and σ_c dependence on ε will change:

$$\sigma_c = 8\pi\left(\frac{Ze^2m^*\varepsilon}{\hbar^2k_0^2}\right)^2. \tag{57.4}$$

This yields for σ in the previous assumptions the value of $\sigma \cong 2 \cdot 10^{-13}$ cm^2.

The expressions obtained above are valid if *the particle being scattered does not affect the field of the scattering centre.* Actually, this is not the case. *The field of the free carrier polarizes the impurity atom inducing a dipole moment.* The value **d** of this moment depends on the intensity of the field established by the scattered particle and on the polarizability χ of the atom:

$$|\mathbf{d}|=\frac{\chi e}{\varepsilon r^3}. \tag{57.5}$$

The induced dipole moment, in turn, establishes an electric field which acts on the free charge. As is known from electrostatics, the intensity of a dipole field in the direction of its axis is

$$E_d=\frac{2d}{\varepsilon r^3}=\frac{2\chi e}{\varepsilon^2 r^5}. \tag{57.6}$$

The potential energy of a free charge carrier in the field of a dipole induced by it will, according to (57.8), be equal to

$$V=-\frac{\chi e^2}{2\varepsilon^2 r^4}. \tag{57.7}$$

The expression (57.7) will be valid insofar as *the distance between the scattering centre and the carrier remains large as compared to* k_0^{-1} since the expression (57.6), valid only for a point dipole, is not applicable at shorter distances. For shorter distances the distribution of the charge of the scattering centre affected by the field of the scattered particle should be found. However, at sufficiently large distances at which the value of the screening potential is practically zero, the charge carriers will be scattered by a neutral scattering centre by means of its *polarization potential* (57.7). Find the differential effective cross section for the scattering by the potential (potential energy) of the form (57.7). To this end calculate $V_\varkappa$:

$$\frac{V_\varkappa}{4\pi}=\int_{R_0}^{\infty} rV(r)\frac{\sin\varkappa r}{\varkappa}\,dr=-\frac{\chi e^2}{2\varepsilon^2}\int_{R_0}^{\infty}\frac{\sin\varkappa r}{\varkappa r^3}\,dr. \tag{57.8}$$

As distinct from the Coulomb field, the expression (57.8) converges for $r\to\infty$, but diverges for $r\to 0$. However, in the last case one can easily get rid of the divergence by introducing a minimum distance R_0 since, as was already stated, the expression (57.7) is not valid for $r\to 0$.

$V_\varkappa$ from (57.8) may be calculated with the aid of complex variables. However, we shall adopt another method Note first of all that for $\varkappa=0$ the integral may be easily calculated:

$$\frac{V_{\varkappa=0}}{4\pi}=-\frac{\chi e^2}{2\varepsilon^2}\int_{R_0}^{\infty}\frac{dr}{r^2}=\frac{\chi e^2}{2\varepsilon^2 R_0}. \tag{57.9}$$

The same expression will hold for $\theta=0$ for all charge carrier energies, or for $k=0$. One may presume the expression (57.9) to be applicable for the case of small kR_0, as well. This leads us to the conclusion that *for $kR_0 \ll 1$ the differential effective scattering cross section by a polarization potential is isotropic and does not depend explicitly on the energy*:

$$\sigma = \frac{4m^{*2}}{\hbar^4} \frac{\chi^2 e^4}{4\varepsilon^4 R_0^2}. \tag{57.10}$$

For large $\varkappa$, $\sin \varkappa r$ is a rapidly oscillating function, therefore the integral (57.9) for large θ is practically zero. But for small θ, when $\frac{\sin \varkappa r}{\varkappa} \cong r$, the integral assumes the value $\frac{e^2 \chi}{2\varepsilon^2 R_0}$. Thus, *scattering by polarized neutral centre is highly anisotropic*. The scattering takes place only inside a cone of an angle $\Delta\theta \cong \frac{1}{kR_0}$, there being practically no scattering outside this cone. The effective scattering cross section inside the cone is little dependent on the energy. However, since the variation of k results in the variation of $\Delta\theta$, the differential effective cross section varies, too, approximately as $(kR_0)^{-2}$. Calculating the transport effective cross section we obtain $\sigma_c \sim (kR_0)^{-4}$. Thus, *as the energy of the scattered particle increases, isotropic scattering turns into a highly anisotropic one*. However, in actual fact things are much more complicated. Since the scattered electron and the atomic electron are identical, scattering may be accompanied by *an exchange effect*. Moreover, it should be kept in mind that the hydrogen-like neutral scattering centre has several excited states in consequence of which the scattering may turn out to be non-elastic. In other words, the problem of the scattering of electrons and holes by neutral centres is much more complicated than the problem of particle scattering by charged impurities. Considering the problem of charge carrier scattering in the field of hydrogen-like impurity atom, C. Erginsoy demonstrated that σ_c is approximately proportional to v^{-1}, and that, consequently, the relaxation time is independent of the energy of the particles being scattered. As a rule, an *approximate relation* obtained by numerical solution methods is used to evaluate the relaxation time:

$$\tau_n = \frac{m^{*2} e^3}{20\varepsilon\hbar^3} \frac{1}{N_n}, \tag{57.11}$$

where N_n is the concentration of neutral impurity at the temperature in question. It should be noted that electroactive impurity can be neutral only at very low temperatures. Note in conclusion of the section that for dislocation scattering the effective scattering cross section depends appreciably on the angle between the original charge carrier velocity and the dislocation axis. σ may be found with the aid of the general equation (55.30).

Omitting computations we would like to remark that the expression for relaxation time in case of edge dislocation scattering currently in use is

$$\tau_d = \frac{3}{8R_d v N_d}, \tag{57.12}$$

where R_d is the radius; N_d, the dislocation density.

Summary of Sec. 57

1. The charge carrier scattering by neutral scattering centres is isotropic. Anisotropy becomes marked only for very high electron and hole energies.

2. The relaxation time in case of charge carrier scattering by impurity atoms or point defects is practically independent of particle velocities. For scattering by edge dislocations $\tau \sim v^{-1}$.

58. LATTICE VIBRATIONS. NORMAL CO-ORDINATES. PHONONS

In order to be able to understand numerous physical phenomena in semiconductors one should have an idea of the methods of describing lattice vibrations. From a general physics course it is known that the atoms of a heated body take part in random thermal vibrations around their equilibrium positions. The amplitude of thermal vibrations depends on the temperature. Electrons and holes, moving in a substance the atoms of which are in a state of random motion, can exchange energy with the atoms. Through thermal vibrations the thermodynamic equilibrium between the lattice and the electron gas is established. This equilibrium is characterized by a definite temperature which is the same both for the lattice atoms (or ions) and for the electron gas. Turn now to the mathematical description of lattice vibrations.

Denote the equilibrium co-ordinate of the atom in site j of an elementary cell, the position of which is described by the vector $\mathbf{n} = (n_1a_1; n_2a_2; n_3a_3)$, by $\mathbf{R}^0_{\mathbf{n}j}$. Suppose the atom $\mathbf{R}^0_{\mathbf{n}j}$ is displaced from its equilibrium position to the point with the co-ordinate $\mathbf{R}'_{\mathbf{n}j}$. The vector

$$\boldsymbol{u}_{\mathbf{n}j} = \mathbf{R}'_{\mathbf{n}j} - \mathbf{R}^0_{\mathbf{n}j} \tag{58.1}$$

is termed displacement vector of the jth atom of the nth cell. Let $U_0 = U_0(\ldots, \mathbf{R}^0_{nj} \ldots,)$ be the equilibrium value of the lattice potential energy, and U the potential energy of the lattice with displaced atoms. The latter may be represented as a function of the atomic displacements:

$$U = U(\ldots, \mathbf{R}_{nj}\ldots,) = U(\ldots, \boldsymbol{u}_{nj}\ldots,). \qquad (58.2)$$

Expand U in the Taylor series in the displacement vector around the equilibrium state of the crystal:

$$U(\ldots, \boldsymbol{u}_{nj}\ldots,) = $$

$$= U_0 + \sum_{nj} \frac{\partial U}{\partial \boldsymbol{u}_{nj}} \boldsymbol{u}_{nj} + \frac{1}{2} \sum_{nj,\, n'j'} \frac{\partial^2 U}{\partial \boldsymbol{u}_{nj}\, \partial \boldsymbol{u}_{n'j'}} \boldsymbol{u}_{nj} \boldsymbol{u}_{n'j'} +$$

$$+ \frac{1}{6} \sum_{nj,\, n'j',\, n''j''} \frac{\partial^3 U}{\partial \boldsymbol{u}_{nj}\, \partial \boldsymbol{u}_{n'j'}\, \partial \boldsymbol{u}_{n''j''}} \boldsymbol{u}_{nj} \boldsymbol{u}_{n'j'} \boldsymbol{u}_{n''j''} + \ldots . \qquad (58.3)$$

Since the expansion is performed around the equilibrium state where the value of the body potential energy is minimum, $\frac{\partial U}{\partial \boldsymbol{u}_{nj}} = 0$, and the series, consequently, starts with the quadratic terms.

Denote the elements of rank II and rank III tensors by

$$A_{nj,\, n'j'} = \frac{\partial^2 U}{\partial \boldsymbol{u}_{nj} \partial \boldsymbol{u}_{n'j'}} = A_{n'j',\, nj}, \qquad (58.4)$$

$$B_{nj,\, n'j',\, n''j''} = \frac{\partial^3 U}{\partial \boldsymbol{u}_{nj} \partial \boldsymbol{u}_{n'j'} \partial \boldsymbol{u}_{n''j''}}, \qquad (58.5)$$

respectively.

Taking into account (58.4) and (58.5) write

$$U = U_0 + \frac{1}{2} \sum_{nj,\, n'} A_{nj,\, n'j'} \boldsymbol{u}_{nj} \boldsymbol{u}_{n'j'} + \ldots . \qquad (58.6)$$

For small displacements, neglecting the cubic terms, represent U as a bi-linear function of atomic displacements. The force acting on the atom **mi** is equal to the derivative of U with respect to $\boldsymbol{u}_{mi}$ with the minus sign:

$$\mathbf{F}_{ml} = -\frac{\partial U}{\partial \boldsymbol{u}_{mi}} = -\frac{1}{2} \sum_{nj,\, n'j'} A_{nj,\, n'j'} (\boldsymbol{u}_{n'j'} \delta_{ml,\, nj} + \boldsymbol{u}_{nj} \delta_{ml.\, nj}) =$$

$$= -\frac{1}{2} \Big(\sum_{n'j'} A_{mi,\, n'j'} \boldsymbol{u}_{n'j'} + \sum_{ni} A_{nj,\, mi} \boldsymbol{u}_{nj} \Big) = -\sum_{nj} A_{mi,\, nj} \boldsymbol{u}_{nj}. \qquad (58.7)$$

The quantity $A_{nj,\, mi}$ depends not on individual values of **m** and **n**, but on the interatomic distance, i.e. on $|\mathbf{n} - \mathbf{m}|$:

$$A_{nj,\, ml} = A_{jl}(\mathbf{n} - \mathbf{m}). \qquad (58.8)$$

It follows from here that a *translation by an integral number of lattice constants does not alter its "elasticity coefficient"*:

$$A_{n+l,\, j;\; m+l,\, l} = A_{jl}(\mathbf{n} - \mathbf{m}) = A_{nj,\, ml}. \qquad (58.9)$$

Write the equation of motion for the mi atom with the mass M_l keeping in mind that $\mathbf{F}_{ml}$ is defined by the relation (58.7):

$$M_l \ddot{\boldsymbol{u}}_{ml} = -\sum_{nj} A_{ml,\,nj} \boldsymbol{u}_{nj}. \tag{58.10}$$

The equation (58.10) explicitly contains M_l, and this in some cases complicates the solution of the problem. Introduce the so-called *reduced displacements* $\mathbf{w}_{nj}$ defined by the condition

$$\mathbf{w}_{nj} = \sqrt{M_j}\, \boldsymbol{u}_{nj}. \tag{58.11}$$

Instead of the matrix $\{A_{nj,\,n'j'}\}$ introduce the matrix $\{D_{nj,\,n'j'}\}$:

$$D_{nj,\,n'j'} = \frac{A_{nj,\,n'j'}}{\sqrt{M_j M_{j'}}}. \tag{58.12}$$

The matrix $\{D_{nj,\,n'j'}\}$ is termed dynamic. The dimensionality of the elements of the D matrix is $[T]^{-2}$. Indeed, since the dimensionality of A coincides with that of the elasticity coefficient

$$[A_{nj,\,ml}] = \left[\frac{F}{L}\right] = \left[\frac{ML}{T^2L}\right] = \left[\frac{M}{T^2}\right], \tag{58.13}$$

the dimensionality of $D_{nj,\,ml}$ coincides with the dimensionality of the square of frequency. It follows from the obvious equality

$$A_{nj,\,ml} \boldsymbol{u}_{nj} \boldsymbol{u}_{ml} = \frac{A_{nj,\,ml}}{\sqrt{M_j M_l}} \sqrt{M_j}\, \boldsymbol{u}_{nj} \sqrt{M_l}\, \boldsymbol{u}_{ml}, \tag{58.14}$$

that the potential energy written in terms of reduced displacements is of the form

$$U = U_0 + \frac{1}{2} \sum_{nj,\,ml} D_{nj,\,ml} \mathbf{w}_{nj} \mathbf{w}_{ml}. \tag{58.15}$$

The equation of motion (58.10) may now be written in the form

$$M_l \ddot{\boldsymbol{u}}_{ml} = \sqrt{M_l}\, \ddot{\mathbf{w}}_{ml} = -\sum_{nj} A_{ml,\,nj} \boldsymbol{u}_{nj} = -\sum_{nj} D_{ml,\,nj} \mathbf{w}_{nj} \sqrt{M_j}, \tag{58.16}$$

or

$$\ddot{\mathbf{w}}_{ml} = -\sum_{nj} D_{ml,\,nj} \mathbf{w}_{nj}. \tag{58.17}$$

The equation systems (58.10) or (58.17) show that *the displacements of different atoms $\boldsymbol{u}_{nj}$* (or $\mathbf{w}_{nj}$) *are interconnected.* A generalized solution of the equation system (58.17) is extremely complicated. In order to simplify the equation system (58.17) we shall use other variables to describe the state of the crystal. Denote the number of elementary cells in the crystal by N, and the number of atoms per elementary cell by s. The number of degrees of freedom of the crystal is equal to $3sN$; they are described by sN vectors $\mathbf{u}_{nj}$, or by their $3sN$ components. *Any set of variables* which may be used to describe the state of the crystal may be introduced instead of $\mathbf{w}_{nj}$. The simplest way to describe the state of the crystal is with the aid of the so-called *normal, or main, co-ordinates.*

Consider with this aim in view *an arbitrary linear transformation* of the co-ordinates (for the sake of simplicity we retain the vector form of the variables).

Suppose we have sN new variables $\mathbf{q}_\alpha$ $(\alpha = 1, \ldots, sN)$ which are connected with the former variables by means of a matrix $\mathbf{S}$ which defines a linear transformation of the co-ordinates of the form

$$\mathbf{q}_\alpha = \sum_{nj} S_{\alpha,\, nj} \mathbf{w}_{nj}. \tag{58.18}$$

Suppose the $\mathbf{S}$ matrix determinant is not equal to zero, and that, therefore, it has a reciprocal matrix which we shall denote by $\mathbf{T}$:

$$\mathbf{T} = \mathbf{S}^{-1};\ \mathbf{ST} = \mathbf{TS} = \mathbf{1}, \tag{58.19}$$

where $\mathbf{1}$ is a unit matrix. The equation (59.18) may be written in the form

$$\begin{gathered}\sum_{nj} S_{\alpha,\, nj} T_{nj,\, \beta} = \delta_{\alpha\beta}, \\ \sum_{\alpha} T_{nj,\, \alpha} S_{\alpha,\, ml} = \delta_{nj,\, ml} = \delta_{nm}\delta_{jl}.\end{gathered} \tag{58.20}$$

Since the choice of the matrix $\mathbf{S}$ is, in general, arbitrary, consider the case of *a unitary** matrix $\mathbf{S}$ for which it may be written

$$T_{nj,\, \beta} = S^*_{\beta,\, nj};\quad S_{\beta,\, nj} = T^*_{nj,\, \beta}, \tag{58.21}$$

$$\sum_{nj} S_{\alpha,\, nj} S^*_{\beta,\, nj} = \delta_{\alpha\beta}, \tag{58.22}$$

$$\sum_{\alpha} S^*_{\alpha,\, nj} S_{\alpha,\, ml} = \delta_{nj,\, ml} = \delta_{nm}\delta_{jl}. \tag{58.23}$$

From (58.18) we may express $\mathbf{w}_{nj}$ in terms of $\mathbf{q}_\alpha$. To this end premultiply (58.18) by $T_{ml,\, \alpha}$ and perform summation over α to obtain

$$\begin{aligned}\sum_{\alpha} T_{ml,\, \alpha}\mathbf{q}_\alpha &= \sum_{\alpha}\sum_{nj} T_{ml,\, \alpha} S_{\alpha,\, nj}\, \mathbf{w}_{nj} = \\ &= \sum_{nj}\Big(\sum_{\alpha} T_{ml,\, \alpha} S_{\alpha,\, nj}\Big)\mathbf{w}_{nj} = \sum_{nj}\delta_{ml,\, nj}\mathbf{w}_{nj} = \mathbf{w}_{ml},\end{aligned} \tag{58.24}$$

* A matrix $\mathbf{S}$ is termed unitary if its reciprocal matrix coincides with the conjugate matrix: $\mathbf{S}^+ = \mathbf{S}^{-1}$. The conjugate matrix is defined by the condition $\mathbf{S}^+ = \tilde{\mathbf{S}}^*$, where the sign (*) means that the matrix elements are complex conjugate, and (~) that the matrix is transposed.

i.e.

$$\mathbf{w}_{mi} = \sum_{\alpha} T_{mi,\,\alpha}\mathbf{q}_{\alpha} = \sum_{\alpha} S^{*}_{\alpha.\ mi}\mathbf{q}_{\alpha}. \tag{58.25}$$

Find the equation of motion in the new variables:

$$\ddot{\mathbf{w}}_{mi} = \sum_{\alpha} T_{mi,\,\alpha}\ddot{\mathbf{q}}_{\alpha} = -\sum_{nj} D_{mi,\,nj}\mathbf{w}_{nj} = -\sum_{nj,\,\beta} D_{mi,\,nj}T_{nj,\,\beta}\,\mathbf{q}_{\beta}. \tag{58.26}$$

Premultiplying the equation (58.26) by $S_{\gamma,\,mi}$ and performing the summation over mi we obtain:
for the left-hand part of the equation

$$\sum_{mi} S_{\gamma,\,mi}\sum_{\alpha} T_{mi,\,\alpha}\ddot{\mathbf{q}}_{\alpha} = \sum_{\alpha}\Big(\sum_{mi} S_{\gamma,\,mi}T_{mi,\,\alpha}\Big)\ddot{\mathbf{q}}_{\alpha} = \sum_{\alpha}\delta_{\gamma\alpha}\ddot{\mathbf{q}}_{\alpha} = \ddot{\mathbf{q}}_{\gamma}, \tag{58.27}$$

and for the right-hand part of the equation

$$-\sum_{mi} S_{\gamma,\,mi}\sum_{nj,\,\beta} D_{mi,\,nj}T_{nj,\,\beta}\,\mathbf{q}_{\beta} = -\sum_{\beta}\Big(\sum_{mi,\,nj} S_{\gamma.\,mi}D_{mi,\,nj}T_{nj,\,\beta}\Big)\mathbf{q}_{\beta} =$$

$$= -\sum_{\beta}(\mathbf{SDT})_{\gamma\beta}\mathbf{q}_{\beta} = -\sum_{\beta}(\mathbf{SDS}^{-1})_{\gamma\beta}\mathbf{q}_{\beta}. \tag{58.28}$$

Taking into acount (58.27) and (58.28) we write

$$\ddot{\mathbf{q}}_{\gamma} = -\sum_{\beta}(\mathbf{SDS}^{-1})_{\gamma\beta}\,\mathbf{q}_{\beta}. \tag{58.29}$$

Since **S** is an arbitrary matrix, choose it so that the *matrix* $\mathbf{SDS}^{-1}$ *would be diagonal*

$$(\mathbf{SDS}^{-1})_{\gamma\beta} = \omega_{\gamma}^{2}\delta_{\gamma\beta}. \tag{58.30}$$

The dimensionality of $(\mathbf{SDS}^{-1})_{\gamma\beta}$ is that of the square of frequency, so that from the standpoint of dimensionality ω_{γ} is a frequency.

We shall demonstrate below that a matrix **S** of this type, in fact, exists. In the mean time suppose that the condition (58.30) may be imposed on the matrix **S**. In this case the equation of motion (58.29) will be simplified:

$$\ddot{\mathbf{q}}_{\gamma} + \omega_{\gamma}^{2}\mathbf{q}_{\gamma} = 0. \tag{58.31}$$

If follows that *the variables* $\mathbf{q}_{\gamma}$ *are independent and vary sinusoidally:*

$$\mathbf{q}_{\gamma}(t) = \mathbf{q}_{\gamma}^{0}e^{-i\omega_{\gamma}t}. \tag{58.32}$$

The vector $\mathbf{q}_{\gamma}^{0}$ combines the amplitude with the initial phase. *The quantities* ω_{γ} *are termed natural frequencies.* The number of natural frequencies of a crystal is sN; therefore, crystal vibrations are characterized by $3sN$ different modes corresponding to the full number of degrees of freedom of the set of sN atoms.

The independent variables satisfying the equation (58.31) *are termed normal, or main, crystal co-ordinates*. Express the energy of the crystal in normal co-ordinates. The potential energy is,

$$U=U_0+\frac{1}{2}\sum_{nj,\,ml} D_{nj,\,ml}\mathrm{w}_{nj}\mathrm{w}_{ml}=$$

$$=U_0+\frac{1}{2}\sum_{nj,\,ml} D_{nj,\,ml}\sum_{\alpha} T_{nj,\,\alpha}\mathrm{q}_{\alpha}\sum_{\beta} T_{ml,\,\beta}\mathrm{q}_{\beta}=$$

$$=U_0+\frac{1}{2}\sum_{\alpha,\,\beta}\Big(\sum_{nj,\,ml} S^*_{\alpha,\,nj}D_{nj,\,ml}T_{ml,\,\beta}\Big)\mathrm{q}_{\alpha}\mathrm{q}_{\beta}\,. \qquad (58.33)$$

The variables u_{nj}, w_{nj} and the matrices **A** and **D** are real. Should the normal co-ordinates be considered real quantities (i.e. should only *the real part* Re $(\mathrm{q}^0_\alpha e^{-i\omega_\alpha t})$ be considered as the normal co-ordinate, and not $\mathrm{q}^0_\alpha e^{-i\omega_\alpha t}$ itself), the matrix S should be real, too:

$$S^*_{\alpha,\,nj}=S_{\alpha,\,nj}. \qquad (58.34)$$

Taking into account the condition of reality (58.34) of the matrix S and the equation (58.30), write the equation (58.33) in a very simple form:

$$U=U_0+\frac{1}{2}\sum_{\alpha}\omega^2_\alpha\mathrm{q}^2_\alpha. \qquad (58.35)$$

Find the kinetic energy T:

$$\mathrm{T}=\sum_{nj}\frac{1}{2}M_j\dot{u}^2_{nj}=\frac{1}{2}\sum_{nj}\dot{\mathrm{w}}^2_{nj}=\frac{1}{2}\sum_{nj}\Big(\sum_{\alpha}T_{nj,\,\alpha}\dot{\mathrm{q}}_{\alpha}\Big)\Big(\sum_{\beta}T_{nj,\,\beta}\dot{\mathrm{q}}_{\beta}\Big)=$$

$$=\frac{1}{2}\sum_{\alpha\beta}\Big(\sum_{nj}T_{nj,\,\alpha}T_{nj,\,\beta}\Big)\dot{\mathrm{q}}_{\alpha}\dot{\mathrm{q}}_{\beta}=\frac{1}{2}\sum_{\alpha}\dot{\mathrm{q}}^2_{\alpha} \qquad (58.36)$$

For the full energy $H=\mathrm{T}+U$ we obtain

$$H=\mathrm{T}+U=\sum_{\alpha}\Big(\frac{1}{2}\dot{\mathrm{q}}^2_{\alpha}+\frac{1}{2}\omega^2_\alpha\mathrm{q}^2_\alpha\Big)+U_0=\sum_{\alpha}H_\alpha+U_0, \qquad (58.37)$$

where

$$H_\alpha=\frac{\dot{\mathrm{q}}^2_\alpha}{2}+\frac{\omega^2_\alpha\mathrm{q}^2_\alpha}{2} \qquad (58.38)$$

is *the Hamilton function corresponding to the αth normal co-ordinate.* The full lattice energy is the sum of the energies of normal vibrations.

The expression (58.38) coincides with the expression for the full energy of a *harmonic oscillator of unit mass* ($M_\alpha=1$ and $\mathrm{p}_\alpha=\dot{\mathrm{q}}_\alpha$):

$$H_\alpha=\frac{\mathrm{p}^2_\alpha}{2}+\frac{\omega^2_\alpha\mathrm{q}^2_\alpha}{2}. \qquad (58.39)$$

In classical physics the energy of a harmonic oscillator may assume arbitrary values. It is proportional to the square of the amplitude of oscillations: $\mathbf{q}_\alpha^{0\prime} = \operatorname{Re} \mathbf{q}_\alpha^0$:

$$\mathbf{q}_\alpha(t) = \mathbf{q}_\alpha^{0\prime} \sin(\omega_\alpha t + \varphi_\alpha); \quad \dot{\mathbf{q}}_\alpha = \omega_\alpha \mathbf{q}_\alpha^{0\prime} \cos(\omega_\alpha t + \varphi_\alpha) \qquad (58.40)$$

and

$$H_\alpha = \frac{\omega_\alpha^2 \mathbf{q}^{0\prime 2}}{2} \cos^2(\omega_\alpha t + \varphi_\alpha) + \frac{\omega_\alpha^2 \mathbf{q}^{0\prime 2}}{2} \sin^2(\omega_\alpha t + \varphi_\alpha) = \frac{\omega_\alpha^2 \mathbf{q}^{0\prime 2}}{2}. \qquad (58.41)$$

In quantum mechanics the Hamilton operator for a harmonic oscillator is of the form

$$\hat{\mathbf{H}}_\alpha = \frac{\left(-i\hbar \frac{d}{d\mathbf{q}_\alpha}\right)^2}{2} + \frac{\omega_\alpha^2 \mathbf{q}_\alpha^2}{2}, \qquad (58\ 42)$$

and the Schrödinger equation

$$\hat{\mathbf{H}}_\alpha \psi_\alpha = E_\alpha \psi_\alpha, \qquad (58.43)$$

where

$$E_\alpha = \hbar\omega_\alpha \left(v_\alpha + \frac{1}{2}\right); \quad v_\alpha = 0, 1, 2, \ldots \qquad (58.44)$$

is the oscillator energy, and ψ_α, its eigenfunction. Since the lattice Hamiltonian is equal to the sum of independent $\hat{\mathbf{H}}_\alpha$,

$$\hat{\mathbf{H}} = \sum_\alpha \hat{\mathbf{H}}_\alpha + U_0, \qquad (58.45)$$

the lattice energy may be represented as a sum of the energies of the harmonic oscillators:

$$E = U_0 + \sum_\alpha E_\alpha = U_0 + \sum_\alpha \hbar\omega_\alpha \left(v_\alpha + \frac{1}{2}\right). \qquad (58.46)$$

The energy of a quantum-mechanical oscillator $E_\alpha = \hbar\omega_\alpha\left(v_\alpha + \frac{1}{2}\right)$ may change by an amount $\Delta E_\alpha = \hbar\omega_\alpha \Delta v_\alpha$. As is known from quantum mechanics, the selection rule for the quantum number of an oscillator is of the form

$$\Delta v_\alpha = \pm 1. \qquad (58.47)$$

If $\Delta v_\alpha = -1$, the lattice will go over into *one of the lower energy states transmitting the energy $\hbar\omega_\alpha$ to the charge carriers* or to the ambient. The energy quantum $\hbar\omega_\alpha$ was termed *lattice vibrations energy quantum, or phonon*. Hence, the transition $\Delta v_\alpha = -1$ may be termed process of *phonon radiation* by the lattice, and the transition $\Delta v_\alpha = +1$, the process of *phonon absorption* by the lattice.

The problem may be approached from another side. We will presume that the vibration energy E_α is localized in the crystal volume in the form of *free quasiparticles, phonons,* their numbers being v_α of each sort. *The phonons constitute the phonon gas.* In this case $\Delta v_\alpha = +1$ means the generation of a phonon, and $\Delta v_\alpha = -1$—the annihilation of a phonon.

The lattice wave function is represented in the form of a product of the wave functions of the oscillators:

$$\Psi = \prod_\alpha \psi_\alpha(\mathbf{q}_\alpha), \tag{58.48}$$

where $\psi_\alpha(\mathbf{q}_\alpha) = \psi_{\alpha 1}\psi_{\alpha 2}\psi_{\alpha 3}$; for example,

$$\psi_{\alpha 1} = e^{-\frac{q_{\alpha 1}^2 \omega_\alpha}{2\hbar}} H_{v_{\alpha 1}}\left(\frac{q_{\alpha 1}}{q_{\alpha 0}}\right), \tag{58.49}$$

where $q_{\alpha 0} = \sqrt{\frac{\hbar}{\omega_\alpha}}$; $H_{v_{\alpha 1}}$ are the Tchebycheff-Hermite polynomials.

To conclude the section we would like to point out that the motion of the atoms described in the adiabatic approximation in normal co-ordinates will be of the form of harmonic oscillations.

Summary of Sec. 58

1. The random motion of the lattice atoms (ions) is described by the displacement vector $\mathbf{u}_{nj}$. Since atomic interaction changes upon displacement, the potential energy, too, depends on the displacements of the atoms:

$$U = U_0 + \frac{1}{2}\sum_{nj,\, n'j'} A_{nj,\, n'j'}\mathbf{u}_{nj}\mathbf{u}_{n'j'} + \ldots . \tag{58.1s}$$

The force acting on the mith atom is equal to

$$\mathbf{F}_{ml} = -\frac{\partial U}{\partial \mathbf{u}_{ml}} = -\sum_{nj} A_{ml,\, nj}\mathbf{u}_{nj}, \tag{59.2s}$$

and the equation of motion for the mith atom is of the form

$$M_l \ddot{\mathbf{u}}_{ml} = -\sum_{nj} A_{ml,\, nj}\mathbf{u}_{nj}. \tag{58.3s}$$

2. The introduction of reduced displacements $\mathbf{w}_{ml} = \sqrt{M_l}\mathbf{u}_{ml}$ and of the dynamic matrix $D_{nj,\, ml} = \frac{A_{nj,\, ml}}{\sqrt{M_j M_l}}$ transforms the equa-

tion (58.3s) into the equation

$$\ddot{\mathbf{w}}_{ml} = -\sum_{nj} D_{ml,\,nj}\mathbf{w}_{nj}. \qquad (58.4s)$$

3. If a unitary matrix $\mathbf{S}$ is assumed to exist which transforms the dynamic matrix into a diagonal matrix

$$(\mathbf{SDS}^{-1})_{\gamma\beta} = \omega_\gamma^2\delta_{\gamma\beta}, \qquad (58.5s)$$

this matrix will enable new variables $\mathbf{q}_\alpha$ to be introduced,

$$\mathbf{q}_\alpha = \sum_{ml} S_{\alpha,\,ml}\mathbf{w}_{ml}, \qquad (58.6s)$$

which satisfy the equation

$$\ddot{\mathbf{q}}_\alpha + \omega_\alpha^2\mathbf{q}_\alpha = 0. \qquad (58.7s)$$

Independent variables varying harmonically with time,

$$\mathbf{q}_\alpha(t) = \mathbf{q}_\alpha^0 e^{-i\omega_\alpha t}, \qquad (58.8s)$$

are termed normal, or main, crystal lattice co-ordinates.

4. The full energy of lattice vibrations expressed in normal co-ordinates constitutes the energy of a set of harmonic oscillators

$$H = \sum_\alpha H_\alpha = \sum_\alpha (\dot{\mathbf{q}}_\alpha^2 + \omega_\alpha^2\mathbf{q}_\alpha^2) \quad (\text{for } U_0 = 0). \qquad (58.9s)$$

In this case the concept of a phonon may be introduced. The phonons are quasiparticles with the energy $\hbar\omega_\alpha$, their numbers being v_α, the lattice energy may be written in the form

$$E = \sum_\alpha \hbar\omega_\alpha\left(v_\alpha + \frac{1}{2}\right), \quad \text{where } v_\alpha = 0, 1, \ldots . \qquad (58.10s)$$

The set of phonons constitutes phonon gas the properties of which adequately describe the properties of lattice vibrations.

59. ACOUSTICAL AND OPTICAL LATTICE VIBRATIONS

Return now to the equation of motion (58.31). We wrote it down in a simple form presuming that there is such a matrix S, with the aid of which the dynamic matrix may be made diagonal:

$$(\mathbf{SDS}^{-1})_{\alpha\beta} = \omega_\alpha^2\delta_{\alpha\beta}, \qquad (59.1)$$

or

$$\sum_{nj,\,ml} S_{\alpha,\,nj}D_{nj,\,ml}T_{ml,\,\beta} = \omega_\alpha^2\delta_{\alpha\beta}. \qquad (59.2)$$

Postmultiply the equation (59.2) by $S_{\beta,kl}$ and perform summation over β:

$$\sum_{\beta}\sum_{nj,\,ml} S_{\alpha,\,nj}D_{nj,\,ml}T_{ml,\,\beta}S_{\beta,\,kl}=\sum_{nj,\,ml} S_{\alpha,\,nj}D_{nj,\,ml}\,\delta_{ml,\,kl}=$$

$$=\sum_{nj} S_{\alpha,\,nj}D_{nj,\,kl}=\sum_{\beta}\omega_{\alpha}^{2}\delta_{\alpha\beta}S_{\beta,\,kl}=\omega_{\alpha}^{2}S_{\alpha,\,kl}, \qquad (59.3)$$

i.e.

$$\sum_{nj} S_{\alpha,\,nj}\left(D_{nj,\,kl}-\omega_{\alpha}^{2}\delta_{nj,\,kl}\right)=0. \qquad (59.4)$$

The system of equations (59.4) may be regarded as one for the elements of the unknown matrix S. This system is homogeneous. For a non-trivial solution to exist the system determinant should be equal to zero:

$$\left|D_{nj,\,kl}-\omega_{\alpha}^{2}\delta_{nj,\,kl}\right|=0. \qquad (59.5)$$

The roots of this equation determine the set of natural frequencies of lattice vibrations and depend only on the elastic properties of the lattice and on the atomic masses.

Consider *the simplest case* as an example:

$$D_{nj,\,ml}=\frac{A_{nj,\,ml}}{\sqrt{M_jM_l}}\,\delta_{nj,\,ml}=\frac{A_{nj,\,nj}}{M_j}\,\delta_{nj,\,ml}. \qquad (59.6)$$

All the non-diagonal elements of this determinant being zeros, the determinant is equal to the product of diagonal elements:

$$\prod_{nj}\left(\frac{A_{nj,\,nj}}{M_j}-\omega_{\alpha}^{2}\right)=0, \qquad (59.6')$$

whence

$$\omega_{\alpha}^{2}=\frac{A_{nj,\,nl}}{M_i}. \qquad (59.7)$$

This is the usual expression for the frequency of natural oscillations of a particle with a mass M_j acted upon by an elastic force $\mathbf{F}_{nj}=-A_{nj,\,nj}\mathbf{u}_{nj}$. Because of the periodic translational symmetry of the crystal, $A_{nj,\,nj}$ are independent of n: $A_{nj,\,nj}=A_{0j}$ and, consequently, the vibration frequencies of j atoms are identical. However, natural frequencies of atoms of different j's are, generally, different. In this simplest case the crystal natural frequency set consists of s frequencies. This justifies the expectation that in a general case, too, there will be s different sets of natural frequencies.

The equation (59.5) may be derived in a somewhat different way. Let us return to the equations of motion (58.10) and (58.17). We shall seek a particular solution to these equations in the form of harmonic vibrations of all the atoms with one frequency:

$$u_{nj}(t) = u^0_{nj} e^{-i\omega t}, \tag{59.8}$$

$$\mathbf{w}_{nj}(t) = \mathbf{w}^0_{nj} e^{-i\omega t}. \tag{59.9}$$

Substitute (59.8) and (59.9) into the aforementioned equations and cancel out $e^{-i\omega t}$ to obtain

$$-\omega M_l u^0_{ml} = -\sum_{nj} A_{ml,\, nj} u^0_{nj}, \tag{59.10}$$

$$-\omega^2 \mathbf{w}^0_{ml} = -\sum_{nj} D_{ml,\, nj} \mathbf{w}^0_{nj}. \tag{59.11}$$

It may be seen that the equation (59.11) exactly coincides with the equation (59.4), i.e. that *the vibration amplitude* $\mathbf{w}^0_{nj}$ *satisfies the same conditions as the transformation matrix* **S**. This means, however, that *the matrix* **S** *may be constructed from the particular solutions having the form of harmonic oscillations.*

Expanding the determinant of the system (59.11) find all possible natural vibration frequencies ω_α of the crystal. The general solution

$$\mathbf{w}_{nj}(t) = \sum_\alpha \mathbf{w}^0_{nj,\,\alpha} e^{-i\omega_\alpha t} \tag{59.12}$$

may be obtained from the particular solutions

$$\mathbf{w}_{nj,\,\alpha}(t) = \mathbf{w}^0_{nj,\,\alpha} e^{-i\omega_\alpha t}. \tag{59.13}$$

The equation (59.12) is analogous to (58.25). In other words, the transformation (58.25) may be regarded as a linear transformation from the "old" to the "new" co-ordinates, or as the presentation of the general solution of the equation of motion in the form of a set of particular solutions made up of harmonic oscillations. This makes the analogy between the equations (59.11) and (59.4) for $\mathbf{w}^0_{nj,\,\alpha}$ and $S_{\alpha,\, nj}$ more vivid.

We shall consider now some properties of the matrix S, or of the particular solutions. Construct a matrix $\mathbf{w}^0$ from $\mathbf{w}^0_{nj,\,\alpha}$ the columns ofwhich represent amplitudes and phases of the oscillations of the same atom corresponding to various frequencies ω_α. The lines represent amplitudes and phases of oscillations of the same frequency, but of different atoms. Substitute the indices $\mathbf{m}+\mathbf{l}$ and $\mathbf{n}+\mathbf{l}$ for $\mathbf{m}$ and $\mathbf{n}$ in equation (59.11) to obtain

$$\omega^2_\alpha \mathbf{w}^0_{\mathbf{m}+\mathbf{l},\, l;\, \alpha} = \sum_{nj} D_{\mathbf{m}+\mathbf{l},\, l;\, \mathbf{n}+\mathbf{l},\, j} \mathbf{w}^0_{\mathbf{n}+\mathbf{l},\, j;\, \alpha}. \tag{59.14}$$

But $D_{m+l,\, i;\, n+l,\, j} = D_{mi,\, nj}$, therefore the equation (59.14) may be presented in the form

$$\omega_\alpha^2 \mathbf{w}^0_{m+l,\, i;\, \alpha} = \sum_{nj} D_{mi,\, nj} \mathbf{w}^0_{n+l,\, j;\, \alpha}. \tag{59.15}$$

It follows from here that if the elements of a column α of the matrix $\mathbf{w}^0$ are a solution of the equation (59.15), the set of quantities $[\mathbf{w}^0_{n+l,\, j;\, \alpha}]_\alpha$ obtained from $[\mathbf{w}^0_{nj,\, \alpha}]_\alpha$ by transposition of the elements by the amount $\mathbf{l}$ (the translation by the vector $\mathbf{l}$) will, too, be a solution of the same equation. But since all the particular solutions have been found, $[\mathbf{w}^0_{n+l,\, j;\, \alpha}]_\alpha$ coincides with one of them to an accuracy of a constant factor:

$$[\mathbf{w}_{n+l,\, j;\, \alpha}]_\alpha = C_\alpha(\mathbf{l})\, [\mathbf{w}^0_{nj,\, \alpha}]_\alpha. \tag{59.16}$$

The quantity $C_\alpha(\mathbf{l})$ has the properties of an exponent:

$$C_\alpha(2\mathbf{l}) = C^2_\alpha(\mathbf{l}), \tag{59.17}$$

which follows from (59.16). Put

$$C_\alpha(\mathbf{l}) = e^{i(\mathbf{K}_\alpha \mathbf{l})}. \tag{59.18}$$

Since $\mathbf{l}$ may be an arbitrary quantity and since the vibration amplitudes must be finite, it follows that $\mathbf{K}_\alpha$ is a real vector, i.e. that the modulus of $C_\alpha(\mathbf{l})$ is unity. This means that the "translation" of the matrix elements by the vector $\mathbf{l}$ does not change the amplitudes of vibration of different atoms, and that only the phases are changed.

The vector $\mathbf{K}_\alpha$ possesses some obvious properties: if the vector $2\pi\mathbf{b}$ is added to it, the condition (59.18) is not violated; therefore, the vector $\mathbf{K}_\alpha$ is determined to the accuracy of the reciprocal lattice vector, and as a result all the different vectors $\mathbf{K}_\alpha$ occupy a limited volume coinciding with the Brillouin zone for the electron wave vector $\mathbf{k}$. Hence, *the values of* $\mathbf{K}_\alpha$ *may be considered in the same space as* $\mathbf{k}$. It is also obvious that the condition (59.13) will remain unaltered if $(-\mathbf{K}_\alpha)$ is taken instead of $\mathbf{K}_\alpha$. These properties of $\mathbf{K}_\alpha$ and of the equation (59.16) enable $\mathbf{w}^0_{nj}$ to be expressed in terms of $\mathbf{w}^0_{0j}$:

$$\mathbf{w}^0_{nj,\, \alpha} = \mathbf{w}^0_{0j,\, \alpha} e^{i(\mathbf{K}_\alpha \mathbf{n})}. \tag{59.19}$$

The quantity $\mathbf{w}^0_{0j,\, \alpha}$ contains the amplitude and the initial phase of all the atoms j vibrating with the frequency ω_α:

$$\mathbf{w}_{nj,\, \alpha}(t) = \mathbf{w}_{0j,\, \alpha} e^{-i[\omega_\alpha t - (\mathbf{K}_\alpha \mathbf{n})]}. \tag{59.20}$$

Should $\mathbf{r}$ be substituted for $\mathbf{n}$, (59.20) would assume the form

$$\mathbf{w}_{rj,\, \alpha}(t) \doteq \mathbf{w}_{j,\, \alpha}(\mathbf{r},\ t) = \mathbf{w}^0_{0j,\, \alpha} e^{-i[\omega_\alpha t - (\mathbf{K}_\alpha \mathbf{r})]}. \tag{59.21}$$

The expression (59.21) *is the equation of a plane harmonic wave with the frequency ω_α propagating in the direction of the wave vector* $\mathbf{K}_\alpha$. The phase velocity of the wave $v_{ph}=\frac{\omega_\alpha}{K_\alpha}$, and the group velocity $v_g=\frac{d\omega_\alpha}{dK_\alpha}$. A correspondence may be established between the plane wave with a frequency ω_α and a quasiparticle with the energy $\hbar\omega_\alpha$ and a momentum $\hbar\mathbf{K}_\alpha$. This particle is the phonon introduced above. Thus, the phonon gas of quasiparticles with the energy $\hbar\omega_\alpha$ and quasimomentum $\hbar\mathbf{K}_\alpha$ corresponds to a set of harmonic waves $(\omega_\alpha, \mathbf{K}_\alpha)$. There should be a definite functional relationship between the frequency and the wave vector:

$$\omega_\alpha=\omega_\alpha(\mathbf{K}_\alpha). \tag{59.22}$$

Re-write the equation (59.21) taking into account the equation (59.19):

$$\sum_{nj} D_{ml,\,nj}\mathbf{w}^0_{nj,\,\alpha}=\sum_{nj} D_{ml,\,nj}\mathbf{w}^0_{0j,\,\alpha}e^{i(\mathbf{K}_\alpha\mathbf{n})}=\omega^2_\alpha\mathbf{w}^0_{0l,\,\alpha}e^{i(\mathbf{K}_\alpha\mathbf{m})}. \tag{59.23}$$

Re-write the equation (59.23) as follows:

$$\sum_j\left[\sum_n D_{ml,\,nj}e^{i(\mathbf{K}_\alpha\mathbf{n})}\right]\mathbf{w}^0_{0j,\,\alpha}-\omega^2_\alpha\mathbf{w}^0_{0l,\,\alpha}e^{i(\mathbf{K}_\alpha\mathbf{m})}=0. \tag{59.24}$$

The equation system (59.24) contains only s unknown vector quantities $\mathbf{w}^0_{0j}$. Multiply (59.24) by $e^{-i(\mathbf{K}_\alpha\mathbf{m})}$. Introduce the notation $\mathbf{n}-\mathbf{m}=\mathbf{l}$ and re-write the sum over $\mathbf{n}$ in (59.24):

$$\sum_{\mathbf{n}} D_{ml,\,nj}e^{i(\mathbf{K}_\alpha n-m)}=\sum_{\mathbf{l}} D_{ml;\,m+l,\,j}e^{i(\mathbf{K}_\alpha\mathbf{l})}=\sum_{\mathbf{l}} D_{lj}(\mathbf{l})e^{i(\mathbf{K}_\alpha\mathbf{l})}=G_{lj}(\mathbf{K}_\alpha) \tag{59.25}$$

after which the equation (59.24) will assume the form

$$\sum_j G_{lj}(\mathbf{K}_\alpha)\mathbf{w}^0_{0j,\,\alpha}-\omega^2_\alpha\mathbf{w}^0_{0l,\,\alpha}=0. \tag{59.26}$$

The solution of a homogeneous equation system exists if the determinant is zero:

$$|G_{lj}(\mathbf{K}_\alpha)-\omega^2_\alpha\delta_{lj}|=0. \tag{59.27}$$

The determinant (59.27) is a $3s$ degree equation with respest to ω^2_α. This means that *there are $3s$, in general, different dependences* $\omega^2_\alpha(\mathbf{K}_\alpha)$, or $3s$ functions $\omega^2_1(\mathbf{K})$, $\omega^2_2(\mathbf{K})$, ..., $\omega^2_{3s}(\mathbf{K})$. The degree of equation (59.5) relative to ω^2 is $3Ns$, and all possible frequencies are determined from this equation, the degree of equation (59.27) with respect to ω^2 is, on the other hand, $3s$; but it serves to determine $3s$ *functional relations* $\omega^2(\mathbf{K})$, and from these for known $\mathbf{K}$'s all the natural frequencies are obtained. Having found all the roots of the equation (59.27) we may solve the equation system (59.26) by substituting into it in turn $\omega_1^2(\mathbf{K}), \ldots\ldots, \omega_{3s}^2(\mathbf{K})$. There is a solution $\mathbf{w}^0_{0j,\alpha}(1), \ldots, \mathbf{w}^0_{0j,\alpha}(s)$ for each of these values. It follows that there are s different values of $\mathbf{w}^0_{0j,\alpha} = \mathbf{w}^0_{0j}$ for each atom j. Expand $\mathbf{w}^0_{0j}$ in three unit vectors one of which ($\mathbf{e}_3$) is collinear with the vector $\mathbf{K}$, and the two others ($\mathbf{e}_1$ and $\mathbf{e}_2$) lie in a plane perpendicular to it. The vibrations along $\mathbf{e}_1$ and $\mathbf{e}_2$ are *transverse*, and along $\mathbf{e}_3$—*longitudinal*. Hence, *there are in a crystal 2s types of transverse waves and s types of longitudinal waves*, 3s wave types in all.

Consider the simplest case: $s=1$. There is only one atom of mass M per elementary cell. The quantity $G_{ij}(\mathbf{K}_\alpha)$ assumes the form

$$G(\mathbf{K}_\alpha) = D(\mathbf{l})\, e^{i(\mathbf{K}_\alpha \mathbf{l})} = \sum_{\mathbf{l}} \frac{A(\mathbf{l})}{M} e^{i(\mathbf{K}_\alpha \mathbf{l})} = \sum_{\mathbf{l}} \frac{A_{\mathbf{n},\,\mathbf{n}+\mathbf{l}}}{M} e^{i(\mathbf{K}_\alpha \mathbf{l})}. \tag{59.28}$$

It follows from (59.27) that

$$\omega_\alpha^2 = G(\mathbf{K}\ \ ) = \sum_{\mathbf{l}} \frac{A_{\mathbf{n},\,\mathbf{n}+\mathbf{l}}}{M} e^{i(\mathbf{K}_\alpha \mathbf{l})}. \tag{59.29}$$

If

$$A_{\mathbf{n},\,\mathbf{n}+\mathbf{l}} = \begin{cases} A_{00} & \text{for } \mathbf{l}=0 \\ 0 & \text{for } \mathbf{l}\neq 0, \end{cases}$$

then

$$\omega_\alpha^2 = \frac{A_{00}}{M}; \quad \omega_\alpha = \sqrt{\frac{A_{00}}{M}}, \tag{59.30}$$

i.e. the frequency ω_α is independent of $\mathbf{K}_\alpha$; therefore $\frac{d\omega_\alpha}{dK_\alpha} = 0$, and the graph $\omega_\alpha(\mathbf{K}_\alpha)$ consists of straight lines parallel to the coordinate axes. The phase velocity in this case will be the smaller the larger $\mathbf{K}_\alpha$, or the shorter the wavelength $\lambda_\alpha = \frac{2\pi}{K_\alpha}$. For $K_\alpha \to 0$ ($\lambda \to \infty$) the phase velocity tends to infinity, therefore all the atoms are displaced as a whole (in case of low frequencies).

Suppose now that besides A_{00} the terms of the A_{01}-type are nonzero. Consider, for the sake of simplicity, a linear chain of atoms (omitting the index α):

$$\begin{aligned} \omega^2 &= \frac{A_{00}}{M} + \frac{A_{01}}{M} e^{iKa} + \frac{A_{01}}{M} e^{-iKa} = \\ &= \omega_0^2 \left[1 + \frac{2A_{01}}{A_{00}} \cos Ka\right] = \omega_0^2 [1 + \gamma \cos Ka], \end{aligned} \tag{59.31}$$

where $\omega_0^2 = \frac{A_{00}}{M}$; $\gamma = \frac{2A_{01}}{A_{00}}$.

There must be a definite relationship between the various A_{ij}. Indeed, suppose all the atoms of the chain are displaced as a whole so that all $u_{nj} = u_0$. In this case

$$\mathbf{F}_{ml} = -\left(\sum_{nj} A_{ml,\,nj} u_{nj}\right) = -\left(\sum_{nj} A_{ml,\,nj}\right) u_0. \qquad (59.32)$$

But since the atomic chain is displaced as a whole, $\mathbf{F}_{ml} = 0$; therefore $\sum\limits_{nj} A_{ml,\,nj} = 0$ and, consequently, in this case

$$A_{00} = -(A_{01} + A_{0,\,-1}). \qquad (59.33)$$

It follows from considerations of symmetry or from the equation

$$u_m = 0; \quad u_{m-1} = u_0; \quad u_{m+1} = -u_0,$$
$$\mathbf{F}_m = (A_{01} - A_{0,\,-1})\, u_0 = 0, \qquad (59.34)$$

that $A_{01} = A_{0,\,-1}$ and $A_{00} = 2A_{01}$, therefore

$$\omega = \omega_0 (1 - \cos Ka)^{1/2} = \omega_0 \sqrt{2} \sin \frac{Ka}{2}. \qquad (59.35)$$

For small K the phase velocity is

$$v_{ph} = \frac{\omega}{K} = \sqrt{2}\omega_0 \frac{\sin \frac{Ka}{2}}{K} \cong \frac{\sqrt{2}}{2} \omega_0 a, \qquad (59.36)$$

and the group velocity

$$v_g = \frac{d\omega}{dK} = \frac{\sqrt{2}}{2} \omega_0 a = v_{ph}, \qquad (59.37)$$

i.e. the group and the phase velocities coincide. A certain material meaning may be attributed to the expression $v_{ph} = v_g$. Denote by $T_s = \frac{A_{00}a}{2}$ the average "strain" of the linear chain, and by $\rho_M = \frac{M}{a}$ the average density of the substance. In this case we may write for the velocity of waves the expression

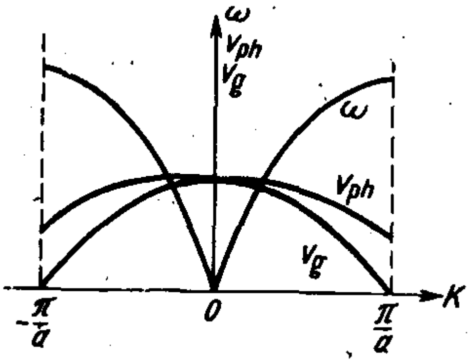


Fig. 82. The dependence of the frequency ω and of the phase and group velocities of the waves in a linear chain on the wave number

$$v_g = v_{ph} = \sqrt{\frac{A_{00}a^2}{2M}} = \sqrt{\frac{T_s}{\rho_M}}. \qquad (59.38)$$

It coincides with the expression for the velocity of elastic waves, or for the speed of sound.

Thus, for $K \longrightarrow 0$ we have the equality $v_g = v_{ph} = c$, where c is the velocity of elastic waves, or the speed of sound. For $K \cong \pm \frac{\pi}{a}$ the group velocity turns zero. Figure 82 shows the dependences $\omega(K)$, $\frac{\omega(K)}{K}$ and $\frac{d\omega(K)}{dK}$ on the wave number for the case discussed above.

Consider now complex cells. Let the cell contain two atoms with the masses M_1 and M_2 and with the co-ordinates $\mathbf{j}_1 = 0$ and $\mathbf{j}_2 = \frac{1}{2}(a, a, a)$. Denote the elasticity coefficient A_{ij} by $A_{11}(\mathbf{l})$; $A_{12}(\mathbf{l}) = A_{21}(\mathbf{l})$ and $A_{22}(\mathbf{l})$ (for an arbitrary $\mathbf{l}$) and the elements of the dynamic matrix by D_{11}, $D_{12} = D_{21}$ and D_{22}. Now we may calculate G_{11}, $G_{21} = G_{12}^*$ and G_{22}. Write, for instance, the equation (59.25) for G_{12}:

$$G_{12}(\mathbf{K}) = \sum_{\mathbf{l}} D_{12}(\mathbf{l})\, e^{i(\mathbf{Kl})}. \qquad (59.39)$$

Write the determinant (59.27)

$$\begin{vmatrix} G_{11} - \omega^2 & G_{12} \\ G_{21} & G_{22} - \omega^2 \end{vmatrix} = 0. \qquad (59.40)$$

From (59.40) find ω^2:

$$(G_{11} - \omega^2)(G_{22} - \omega^2) - G_{12}G_{21} = 0, \qquad (59.41)$$

or

$$\omega^4 - \omega^2 (G_{11} + G_{22}) + G_{11}G_{22} - G_{12}G_{21} = 0. \qquad (59.42)$$

Solving the equation (59.42) we obtain

$$\omega^2 = \frac{G_{11} + G_{22}}{2} \pm \sqrt{\frac{(G_{11} + G_{22})^2}{4} - G_{11}G_{22} + G_{12}G_{21}} = $$
$$= \frac{G_{11} + G_{22}}{2} \pm \sqrt{\frac{(G_{11} - G_{22})^2}{4} + G_{12}G_{21}}, \qquad (59.43)$$

i.e. ω^2 assumes two different values for each $\mathbf{K}$.

Let, for example, $G_{11} = G_{22}$. Then

$$\omega_1^2 = G_{11} + |G_{12}|, \qquad (59.44)$$
$$\omega_2^2 = G_{11} - |G_{12}|. \qquad (59.45)$$

Re-write the system of equations (59.26) in the following form, omitting the subscripts 0 and α of $\mathbf{w}_{0\mathbf{j},\alpha}^0$:

$$(G_{11} - \omega^2)\mathbf{w}_1^0 + G_{12}\mathbf{w}_2^0 = 0, \qquad (59.46)$$
$$G_{21}\mathbf{w}_1^0 + (G_{22} - \omega^2)\mathbf{w}_2^0 = 0.$$

Since there are two different values ω_1^2 and ω_2^2 of ω^2, substituting them in turn into (59.46), we obtain two systems of equations: for $\omega_1^2 = G_{11} + |G_{12}|$

$$\left.\begin{aligned} -|G_{12}|\,\mathbf{w}_1^0 + G_{12}\mathbf{w}_2^0 = 0, \\ G_{21}\mathbf{w}_1^0 - |G_{12}|\,\mathbf{w}_2^0 = 0; \end{aligned}\right\} \tag{59.47}$$

for $\omega_2^2 = G_{11} - |G_{12}|$

$$\left.\begin{aligned} |G_{12}|\,\mathbf{w}_1^0 + G_{12}\mathbf{w}_2^0 = 0, \\ G_{12}\mathbf{w}_1^0 + |G_{12}|\,\mathbf{w}_2^0 = 0. \end{aligned}\right\} \tag{59.48}$$

It follows from (59.47) that if $\omega^2 = G_{11} + |G_{12}|$, then

$$\mathbf{w}_1^0 = + \frac{G_{12}}{|G_{12}|}\mathbf{w}_2^0. \tag{59.49}$$

In the same way it follows from (59.48) that for $\omega^2 = G_{11} - |G_{12}|$

$$\mathbf{w}_1^0 = - \frac{G_{12}}{|G_{12}|}\mathbf{w}_2^0. \tag{59.50}$$

Suppose that G_{12} is real and positive. In this case we will have vibrations of two types:

$$\mathbf{w}_1^0 = \mathbf{w}_2^0 \text{ (type } A), \tag{59.51}$$

$$\mathbf{w}_1^0 = - \mathbf{w}_2^0 \text{ (type } O). \tag{59.52}$$

In the A-type vibrations *the displacements of both atoms are the same, and the cell is displaced as a whole.* This results in local compressions and extensions in the crystal analogous to those which take place when elastic, or acoustical, waves propagate through the crystal. For this reason *the vibrations in the course of which both atoms vibrate in the same phase are termed acoustical.*

In the O-type vibrations *different atoms are displaced in opposite directions, they vibrate in antiphase. The centre of masses of the cell remains stationary, but the centres of summary charges of opposite signs are displaced, and because of this a dipole electric moment appears in the cell.* The vibrations of the O-type were termed *optical vibrations* because they interact strongly with light (with electromagnetic waves) passing through the solid.

Generally, G_{12} may be a complex quantity, therefore, represent it in the form

$$G_{12} = |G_{12}|\,e^{i\varphi}, \tag{59.53}$$

after which the equations (59.49) and (59.50) will assume the form

$$\mathbf{w}_1^0 = \mathbf{w}_2^0 e^{i\varphi} \tag{59.54}$$

and

$$\mathbf{w}_1^0 = -e^{i\varphi}\mathbf{w}_2^0 = e^{i(\varphi+\pi)}\mathbf{w}_2^0, \tag{59.55}$$

respectively.

For $\varphi = \frac{\pi}{2}$ the difference between A- and O-type vibrations vanishes since in both cases the phase shift is equal to $\pm\frac{\pi}{2}$. The closer is φ to 0 or to π, the more pronounced is the difference between acoustical and optical lattice vibrations, i.e. the greater is the change of the distribution of charge and mass in the cell.

Consider a case of a more general nature, when $G_{11} \neq G_{22}$. Substituting into (59.46) the value of ω^2 from (59.43) we obtain two equations connecting $\mathbf{w}_1^0$ and $\mathbf{w}_2^0$:

$$\mathbf{w}_2^{0\,(1)} = \left\{\frac{G_{11}+G_{22}}{2} + \sqrt{\frac{(G_{11}-G_{22})^2}{4} + |G_{12}|^2}\right\}\frac{1}{G_{12}}\mathbf{w}_1^{0\,(1)} \tag{59.56}$$

and

$$\mathbf{w}_2^{0\,(2)} = \left\{\frac{G_{11}+G_{22}}{2} - \sqrt{\frac{(G_{11}-G_{22})}{4} + |G_{12}|^2}\right\}\frac{1}{G_{12}}\mathbf{w}_1^{0\,(2)}, \tag{59.57}$$

which correspond to the plus and minus signs in front of the radical in (59.43). As in the special case of $G_{22} = G_{11}$, one solution corresponds to the "in-phase" vibration of the atoms of the cell, the second solution corresponding to the "anti-phase" vibration; in other words, the solutions correspond to the acoustical and the optical branches, respectively. Consider the dependence $\omega(\mathbf{K})$ for the optical and the acoustical branch in the range of small K (long waves).

To obtain the dependence $\omega(\mathbf{K})$ one should, when calculating G_{ij}, take the terms with $\mathbf{l} \neq 0$ in addition to the terms $\mathbf{l} = 0$. Consider, for the sake of simplicity, a unidimensional atomic chain. We shall confine ourselves to the simplest case: we shall take into account only the interaction between the nearest atoms, i.e. we shall suppose $A_{11}(0)$; $A_{22}(0)$; $A_{12}(0)$; $A_{21}(0)$; $A_{12}(\mathbf{l})$, $A_{21}(\mathbf{l})$ to be non-zero. For the sake of simplicity put $A_{11}(0) = A_{22}(0) = A$; $A_{12}(0) = A_{21}(0) = A_{12}(\mathbf{l}) = A_{21}(\mathbf{l}) = B$. There is a definite relationship between A and B; since

$$\sum_{\mathbf{l}j} A_{\mathbf{l}j}(\mathbf{l}) = A_{11}(0) + A_{12}(0) + A_{12}(\mathbf{l}) = A + 2B = 0 \tag{59.58}$$

it follows

$$A = -2B. \tag{59.59}$$

Calculate G_{ij}:

$$\left.\begin{aligned}
G_{11} &= \sum_{\mathbf{l}} D_{11} e^{i(\mathbf{Kl})} = \frac{A_{11}(0)}{M_1} = \frac{A}{M_1}; \quad G_{22} = \frac{A}{M_2}, \\
G_{12} &= \frac{A_{12}(0)}{\sqrt{M_1 M_2}} + \frac{A_{12}(\mathbf{l})}{\sqrt{M_1 M_2}} e^{i(\mathbf{Ka})} = \frac{B}{\sqrt{M_1 M_2}} \left(1 + e^{i(\mathbf{Ka})}\right), \\
G_{21} &= \frac{B}{\sqrt{M_1 M_2}} \left(1 + e^{-i(\mathbf{Ka})}\right) = G_{12}^{*}.
\end{aligned}\right\} \qquad (59.60)$$

The last relation takes account of the fact that the sign of the vector $\mathbf{l}$ directed from the first to the second atom is opposite to that of the vector directed from the second atom to the first.

Substituting (59.60) into (59.43) we obtain:

$$\begin{aligned}
\omega^2 &= \frac{1}{2}\left(\frac{A}{M_1} + \frac{A}{M_2}\right) \pm \sqrt{\left(\frac{A}{M_1} - \frac{A}{M_2}\right)^2 \frac{1}{4} + \frac{B^2}{M_1 M_2} \left|1 + e^{i(\mathbf{Ka})}\right|^2} = \\
&= \frac{A}{2}\left(\frac{M_1 + M_2}{M_1 M_2}\right)\left[1 \pm \sqrt{1 - \frac{4 M_1 M_2}{(M_1 + M_2)^2} \sin^2 \frac{Ka}{2}}\right] = \\
&= \omega_0^2 \left[1 \pm \sqrt{1 - \gamma^2 \sin^2 \frac{Ka}{2}}\right], \qquad (59.61)
\end{aligned}$$

where

$$\omega_0^2 = \frac{A}{2}\left(\frac{M_1 + M_2}{M_1 M_2}\right); \quad \gamma^2 = \frac{4 M_1 M_2}{(M_1 + M_2)^2} \leqslant 1; \quad \omega_0^2 \gamma^2 = \frac{2A}{M_1 M_2}.$$

Consider the dependence $\omega(K)$ in the vicinity of $K = 0$. Setting $\sin \frac{Ka}{2} \cong \frac{Ka}{2}$ we obtain

$$\omega^2 = \omega_0^2 \left[1 \pm \left(1 - \gamma^2 \frac{K^2 a^2}{4}\right)\right]. \qquad (59.62)$$

For the branch $\omega(K)$ corresponding to the minus sign in (59.62) we have

$$\omega(K) = \frac{\omega_0 \gamma K a}{2} = \sqrt{\frac{A}{2(M_1 + M_2)}}\, Ka. \qquad (59.63)$$

The phase v_{ph} and group v_g velocities are equal and constitute the speed of sound:

$$v_{ph} = v_g = \frac{\omega}{K} = \frac{d\omega}{dK} = \frac{\omega_0 \gamma a}{2} = \sqrt{\frac{A a^2}{2(M_1 + M_2)}} = c. \qquad (59.64)$$

Substituting $\omega = 0$ and $K = 0$ into (59.46) we obtain

$$\frac{\mathbf{w}_2^0}{\mathbf{w}_1^0} = \sqrt{\frac{M_2}{M_1}}; \quad \frac{u_1^0}{u_2^0} = 1, \qquad (59.65)$$

i.e. the atoms vibrate in phase with the same amplitude. In other words, the branch corresponding to the minus sign in front of the radical (59.61) describes acoustical vibrations. The plus sign in (59.61) describes optical vibrations:

$$\omega(K)=\sqrt{2}\omega_0\left(1-\frac{\gamma^2a^2K^2}{16}\right). \qquad (59.66)$$

For $K=0$ the frequencies $\omega(0)=\sqrt{2}\omega_0=\sqrt{A\left(\frac{M_1+M_2}{M_1M_2}\right)}$. With the increase in K the frequency of the optical vibrations ω decreases in accordance with a quadratic law. For $K\rightarrow 0$, the phase velocity of optical vibrations tends to infinity, and for $K=0$ their group velocity becomes zero. The minimum frequency of optical vibrations corresponds to $\frac{aK}{2}=\frac{\pi}{2}$ and is equal to

$$\omega^O\left(\frac{\pi}{a}\right)=\omega_0\sqrt{1+\sqrt{1-\gamma^2}}=\omega^O_{min}. \qquad (59.67)$$

The same condition applies to the maximum frequency of acoustical vibrations which is equal to

$$\omega^A\left(\frac{\pi}{a}\right)=\omega_0\sqrt{1-\sqrt{1-\gamma^2}}=\omega^A_{max}. \qquad (59.68)$$

As may be seen from the general expression (59.61), $\omega^O(K)>$ $>\omega^A(K)$. The frequency of acoustical vibrations varies between zero and ω^A_{max}; the frequency of optical vibrations varies inside the range from $\sqrt{2}\omega_0$ to $\omega_0\sqrt{1+\sqrt{1-\gamma^2}}$. The closer is γ to unity; the greater is the spectral range of the optical and the acoustical vibrations. The group velocity $\frac{d\omega}{dK}$ at point $K=\pm\frac{\pi}{a}$ is zero. For $\gamma=1$ the spectra of the optical and acoustical vibrations merge. For $\gamma<1$ there is a frequency interval between them.

For many computations it may be assumed that there is only one optical vibration frequency $\cong\omega_0$ *which for most crystals lies in the infrared range of the radiation spectrum.*

Figure 83 shows the graph of the dependence $\omega(K)$ for the optical and acoustical branches.

The properties of the optical and acoustical branches obtained by us are independent of the specific lattice type which affects only the functional dependence of $\omega(K)$. The greatest maximum frequencies of the optical vibrations correspond to longer waves since the phase velocity for the optical branch becomes infinite when $K\rightarrow 0$, while the group velocity in this case turns zero. In case there is no maximum wavelength of the optical and the acoustical branch ($\lambda_{max}=\infty$) there is a minimum wavelength

$$\lambda_{min}=\frac{2\pi}{K_{max}}=\frac{2\pi}{\frac{\pi}{a}}=2a, \qquad (59.69)$$

the meaning of which is quite obvious. Evidently, the minimum distance in space between two points, the difference of vibration phases between which is 2π, for a discrete arrangement of vibrating centres cannot be less than $2a$. The meaning of the discreteness of K, which follows from the periodic boundary conditions, is the same: only such vibrations can take place the wavelengths of which are equal to an integral number of lattice constants.

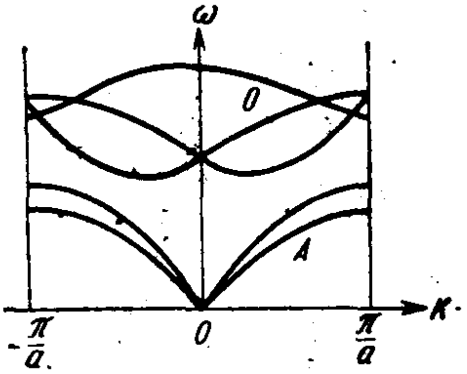


Fig. 83. The optical and acoustical branches of lattice vibrations

The lattice vibration frequencies lie in the range of 0 to 10^{13} Hz. *Long-wave optical lattice vibrations* (vibrations of the highest frequencies) are observed in crystal absorption and reflection spectra in the range of several tens of microns.

Summary of Sec. 59

1. Solving the equation of motion (59.8) in the assumption that the particular solution is of the form of harmonic oscillations with the frequency ω we obtain an equation for the amplitude coinciding with the equation for the co-ordinate transformation matrix **S**. It follows from the properties of the matrices **S** and **D** that neighbouring atoms vibrate with a constant phase difference. In other words, harmonic vibrations constitute travelling harmonic waves in the direction of the vector **K**.

2. Introducing the matrix **G**, whose elements are the elements of the dynamic matrix multiplied by some phase factor, we reduce the system of $3sN$ equations to that of $3s$ equations, the solution of which yields $3s$ different branches of the ω(**K**) dependence. Three of these branches are termed acoustical, and the other $(3s-3)$—optical.

3. The acoustical branches correspond to in-phase vibration of different atoms belonging to the same elementary cell, when all the atoms of the cell vibrate with the same amplitude and in the

same phase. The relationship between the frequency and the wave number in the vicinity of the centre of Brillouin zone is linear:

$$\omega^A = cK, \tag{59.1s}$$

where c is the speed of sound.

4. Atomic vibrations corresponding to the optical branch are anti-phase vibrations with the amplitudes inversely proportional to atomic masses. As a result the centre of masses of the cell remains stationary, while the atomic displacements result in the separation of charges thereby creating an electric dipole moment in the cell.

5. The difference between the acoustical and optical vibrations is apparent only in the vicinity of $\mathbf{K}=0$. Away from this point the difference between the optical and acoustical vibrations disappears.

60. LATTICE SPECIFIC HEAT. PHONON STATISTICS

Statistical physics contains a proof of the statement that *the probability w_s for a system to be in a state with the energy E_s is equal to*

$$w_s = \frac{1}{Z} e^{-\frac{E_s}{kT}}, \tag{60.1}$$

where *the statistical sum* Z is determined by the normalizing condition

$$\sum_s w_s = 1, \tag{60.2}$$

whence the expression for Z is

$$Z = \sum_s e^{-\frac{E_s}{kT}}. \tag{60.3}$$

The average, or the equilibrium, energy value $\langle E\rangle$ is given by formula

$$\langle E\rangle = \sum_s E_s w_s. \tag{60.4}$$

Make use of the expression (60.4) to find the equilibrium energy value for a harmonic oscillator:

$$\langle E_\alpha\rangle = \frac{\sum\limits_{n=0}^{\infty} \hbar\omega_\alpha (n+1/2)\, e^{-\frac{\hbar\omega_\alpha (n+1/2)}{kT}}}{\sum\limits_{n=0}^{\infty} e^{-\frac{\hbar\omega_\alpha (n+1/2)}{kT}}} = \frac{\hbar\omega_\alpha}{2} + \frac{\sum\limits_{n=0}^{\infty} n\hbar\omega_\alpha\, e^{-\frac{n\hbar\omega_\alpha}{kT}}}{\sum\limits_{n=0}^{\infty} e^{-\frac{n\hbar\omega_\alpha}{kT}}}. \tag{60.4'}$$

The denominator of (60.4′) may be calculated as the sum of a geometrical progression:

$$\sum_{n=0}^{\infty}\left[e^{-\frac{\hbar\omega_\alpha}{kT}}\right]^n=\frac{1}{1-e^{-\frac{\hbar\omega_\alpha}{kT}}}. \qquad (60.5)$$

However, since

$$\sum_{n=0}^{\infty} nq^n=q\sum_{n=0}^{\infty} nq^{n-1}=q\sum_{n=0}^{\infty}\frac{d}{dq}q^n=q\frac{d}{dq}\frac{1}{(1-q)}=\frac{q}{(1-q)^2}, \qquad (60.6)$$

it follows tnat

$$\sum_{n=1}^{\infty} n\hbar\omega_\alpha e^{-\frac{n\hbar\omega_\alpha}{kT}}=\hbar\omega_\alpha\frac{e^{-\frac{\hbar\omega_\alpha}{kT}}}{\left(1-e^{-\frac{\hbar\omega_\alpha}{kT}}\right)^2}. \qquad (60.7)$$

Taking into account (60.4-60.7) write the expression for the equilibrium energy of the oscillator:

$$\langle E_\alpha\rangle=\frac{\hbar\omega_\alpha}{2}+\frac{\hbar\omega_\alpha e^{-\frac{\hbar\omega_\alpha}{kT}}}{1-e^{-\frac{\hbar\omega_\alpha}{kT}}}=\frac{\hbar\omega_\alpha}{2}+\frac{\hbar\omega_\alpha}{e^{\frac{\hbar\omega_\alpha}{kT}}-1}, \qquad (60.8)$$

which is in accordance with the Bose-Einstein distribution function. It follows from (60.8) that

$$\langle E_\alpha\rangle=\hbar\omega_\alpha\left(\frac{1}{2}+\frac{1}{e^{\frac{\hbar\omega_\alpha}{kT}}-1}\right). \qquad (60.9)$$

Denote the ratio of $\langle E_\alpha\rangle$ to $\hbar\omega_\alpha$ by $\langle n_\alpha\rangle$:

$$\langle n_\alpha\rangle=\frac{\langle E_\alpha\rangle}{\hbar\omega_\alpha}=\frac{1}{2}+\frac{1}{e^{\frac{\hbar\omega_\alpha}{kT}}-1} \qquad (60.10)$$

where $\langle n_\alpha\rangle$ is *the "mean" value of the oscillator quantum number.*

For $T\to\infty$, when $\hbar\omega_0\ll kT$, the oscillator is in a high energy state:

$$\langle n_\alpha\rangle=\frac{1}{2}+\frac{1}{\frac{\hbar\omega_\alpha}{kT}}\cong\frac{kT}{\hbar\omega_\alpha}, \qquad (60.11)$$

and its average energy is equal to kT:

$$\langle E_\alpha\rangle=\langle n_\alpha\rangle \hbar\omega_\alpha=kT, \tag{60.12}$$

in full accordance with classical statistics.

At low temperatures, when $\hbar\omega_\alpha \gg kT$

$$\langle n_\alpha\rangle=\frac{1}{2}+e^{-\frac{\hbar\omega_\alpha}{kT}}\cong\frac{1}{2}, \tag{60.13}$$

i.e. the oscillator is in the lowest state: $n\cong 0$.

Since the zero energy of the oscillator plays no part in the energy balance in the course of the oscillator interaction with the ambient, we shall henceforth neglect it and assume *the average energy of the oscillator to be equal to*

$$\frac{\hbar\omega_\alpha}{e^{\frac{\hbar\omega_\alpha}{kT}}-1}. \tag{60.14}$$

Making use of the results obtained for the harmonic oscillator find the average energy of lattice vibrations. The general expression (60.4) may be used as the starting point, and account should be taken of the fact that

$$E_s=E(v_1, v_2, \ldots)=\sum_\alpha \hbar\omega_\alpha\left(v_\alpha+\frac{1}{2}\right)$$
$$(v_\alpha=0, 1, 2, \ldots). \tag{60.15}$$

Substituting (60.15) into (60.4) we shall obtain the general expression for $\langle E\rangle$ which, because all the oscillators are independent, may be transformed into a sum of average energy values of each of the oscillators (60.10) or (60.14) so that

$$\langle E\rangle=\sum_\alpha\langle E_\alpha\rangle\doteq\sum_\alpha\frac{\hbar\omega_\alpha}{e^{\frac{\hbar\omega_\alpha}{kT}}-1}\sum_{j=1}^{3s}\sum_{\mathbf{K}_\alpha}\frac{\hbar\omega_j(\mathbf{K}_\alpha)}{e^{\frac{\hbar\omega_j(\mathbf{K}_\alpha)}{kT}}-1}. \tag{60.16}$$

The sum (60.16) may be conveniently subdivided into two sums: the sum over the optical ($j>3$) and the sum over the acoustical ($j\leqslant 3$) vibration branches:

$$\langle E\rangle=\sum_{j=1}^{3}\sum_{\mathbf{K}_\alpha}\frac{\hbar\omega_j(\mathbf{K}_\alpha)}{e^{\frac{\hbar\omega_j(\mathbf{K}_\alpha)}{kT}}-1}+\sum_{j=4}^{3s}\sum_{\mathbf{K}_\alpha}\frac{\hbar\omega_j(\mathbf{K}_\alpha)}{e^{\frac{\hbar\omega_j(\mathbf{K}_\alpha)}{kT}}-1}. \tag{60.17}$$

The frequencies of optical vibrations lie within a narrow range, therefore $\omega_j(\mathbf{K}_\alpha) \cong \omega_j^0$, and

$$\sum_{j=4}^{3s} \sum_{\mathbf{K}_\alpha} \frac{\hbar\omega_j^0}{e^{\frac{\hbar\omega_j^0}{kT}}-1} \cong \sum_{j=4}^{3s} \frac{N\hbar\omega_j^0}{e^{\frac{\hbar\omega_j^0}{kT}}-1}, \tag{60.18}$$

since $\mathbf{K}_\alpha$ assumes N different values.

Introduce a characteristic parameter—*the Debye temperature* Θ_j^0—by the relation

$$\hbar\omega_j^0 = k\Theta_j^0. \tag{60.19}$$

Should we put $\omega_j^0 = 1.4 \times 10^{13}\,\mathrm{s}^{-1}$ we would obtain for the Debye temperature $\Theta_j^0 = 100\,\mathrm{K}$. For most substances the values of Θ_j^0 lie in the range of 100-400 K. Using the quantities Θ_j^0 the energy of the optical vibrations may be presented in the form

$$\langle E^0 \rangle = N \sum_{j=4}^{3s} \frac{\Theta_j^0}{e^{\frac{\Theta_j^0}{T}}-1} \tag{60.20}$$

For $T \gg \Theta_j^0$, $e^{\frac{\Theta_j^0}{T}} = 1 + \frac{\Theta_j^0}{T}$ and

$$\langle E^0 \rangle = N \sum_{j=4}^{3s} kT = (3s-3)\, NkT, \tag{60.21}$$

i.e. all the optical vibrations are excited to high quantum states. The quantity $\langle n_\alpha^0 \rangle = \frac{\langle E_\alpha^0 \rangle}{\hbar\omega_\alpha^0}$ may be regarded as *the number of optical phonons*. At $T \gg \Theta_j^0$ there is *a great number of optical phonons* in the crystal, at $T \ll \Theta_j^0$ the optical vibrations are practically absent ($\langle n_\alpha^0 \rangle \cong 0$) and, consequently, do not contribute to the energy of lattice vibrations.

The variation of the frequency in the acoustical branch is quasi-continuous from zero to ω_0^A. In this case the sum over $\omega_j(K_\alpha)$ may be replaced by an integral over the frequencies. However, this requires the knowledge of the density of states over the frequencies, since in a unit frequency interval ($d\omega = 1$) there may be a different number of vibrations. Let the number of vibrations per unit frequency interval be $g(\omega)$. In this case in an interval $d\omega$ there will be

$$dN = g(\omega)\, d\omega \tag{60.22}$$

of them.

To find the function $g(\omega)$ the dispersion relation $\omega(\mathbf{K})$ should be available. If V is the crystal volume, the volume of the phase space $V\Delta\tau_K\hbar^3$ will contain $\frac{\hbar^3\Delta\tau_K V}{h^3}=\frac{V\Delta\tau_K}{8\pi^3}$ elementary cells, one possible vibration state corresponding to each cell. Construct a spherical layer of the radius K and the thickness dK in the **K**-space; the volume of the layer will be $4\pi K^2 dK$, and the corresponding number of cells $\frac{V4\pi K^2 dK}{8\pi^3}$, $\frac{4\pi K^2 dK}{8\pi^3}$ cells per unit crystal volume.

Express this quantity in terms of the frequency. Since $v_g=\frac{d\omega}{dK}$ it follows that $dK=\frac{d\omega}{v_g}$. Expressing K in terms of ω we obtain

$$dN=\frac{K^2(\omega)\,d\omega}{2\pi^2 v_g}=g(\omega)\,d\omega, \qquad (60.23)$$

since $K=K(\omega)$.

Hence, *the density of lattice vibration states* is of the form

$$g(\omega)=\frac{K^2\omega}{2\pi^2 v_g}. \qquad (60.24)$$

In the vicinity of the Brillouin zone centre the phase and group velocities of the acoustical branch are equal to the speed of sound

$$\frac{\omega}{K}=c; \ \frac{d\omega}{dK}=c; \ K=\frac{\omega}{c}, \qquad (60.25)$$

therefore

$$g(\omega)=\frac{\omega^2}{2\pi^2 c^3}. \qquad (60.26)$$

Thus, for the states in the vicinity of the centre of the Brillouin zone, i.e. for long waves, the number of vibrations per unit frequency interval per unit crystal volume is determined by the relation (60.26). *Following Debye, presume that the relation* (60.26) *is valid for all the acoustical vibration spectrum.* Besides, take into account that *two types of transverse and one type of longitudinal vibrations,* with the velocities c_t and c_l, respectively, are possible. In this case the density of states $g(\omega)$ will be

$$g(\omega)=\frac{\omega^2}{2\pi^2}\left(\frac{2}{c_t^3}+\frac{1}{c_l^3}\right)=\frac{3\omega^2}{3\pi^2 c_0^3}, \qquad (60.27)$$

where c_0 determined by the condition

$$\frac{1}{c_0^3}=\frac{1}{3}\left(\frac{2}{c_t^3}+\frac{1}{c_l^3}\right), \qquad (60.28)$$

is *the speed of sound averaged over directions and types of vibrations*. Now write out the energy of acoustical vibrations:

$$\langle E^A\rangle = \int\limits_0^{\omega_{max}^A} \frac{\hbar\omega g(\omega)\, V\, d\omega}{e^{\frac{\hbar\omega}{kT}}-1} = \frac{3V\hbar}{2\pi^2 c_0^3}\int\limits_0^{\omega_{max}^A} \frac{\omega^3\, d\omega}{e^{\frac{\hbar\omega}{kT}}-1}. \qquad (60.29)$$

Put $\frac{\hbar\omega}{kT}=x$ and re-write (60.29) in the following form:

$$\langle E^A\rangle = \frac{3V(kT)^4}{2\pi^2 c_0^3\hbar^3}\int\limits_0^{x_M}\frac{x^3\,dx}{e^x-1}. \qquad (60.30)$$

The quantity x_M, or the corresponding maximum frequency ω_M, may be found from the condition of normalization of $g(\omega)$ to the total number of acoustical vibrations equal to $3N$:

$$\int\limits_0^{\omega_{max}^A} g(\omega)\, d\omega = \frac{3N}{V}. \qquad (60.31)$$

Substituting the equation (60.27) into (60.31) we obtain

$$\int\limits_0^{\omega_{max}^A}\frac{3\omega^2}{2\pi^2 c_0^3}\, d\omega = \frac{\omega_{max}^3}{2\pi^2 c_0^3} = \frac{3N}{V} = \frac{3}{V_a},$$

$$\omega_{max} = \omega_{max}^A, \qquad (60.32)$$

or

$$\omega_{max} = \frac{(6\pi^2)^{1/3} c_0}{V_a} = \frac{(6\pi^2)^{1/3} c_0}{a}, \qquad (60.33)$$

where a is the lattice constant, and $V_a = a^3 = \frac{V}{N}$ is the volume of an elementary cell.

Assess the maximum value of the wave vector:

$$K_{max} = \frac{\omega_{max}}{c_0} = \frac{(6\pi^2)^{1/3}}{a} = \frac{7.67}{a};$$

$$K_{max} \cong 10^8\ \text{cm}^{-1}. \qquad (60.34)$$

Should K_{max} be assessed on the basis of the minimum wavelength $\lambda_{min} = 2a$, we would obtain the value $K_{max} = \frac{\pi}{a}$. Thus, the Debye interpolation results in an overestimate of the maximum value of the lattice wave vector, since in this case *the Brillouin zone is replaced by a sphere of the radius K_{max}.*

Assess the maximum frequency ω_{max}:

$$\omega_{max}=c_0K_{max}; \quad \omega_{max}\cong 10^5\times 10^8=10^{13}\,\mathrm{s}^{-1}. \tag{60.35}$$

Instead of the maximum frequency ω_{max} the acoustical vibration spectrum may be described by *the characteristic temperature of the solid, or by the Debye temperature* Θ:

$$k\Theta=h\omega_{max}=h\omega^A_{max}; \quad \Theta=\frac{\hbar}{k}\left(\frac{6\pi^2}{a^3}\right)^{1/3}c_0. \tag{60.36}$$

For $\omega_{max}\cong 10^{13}\,\mathrm{s}^{-1}$ the Debye temperature is $\Theta=100\,\mathrm{K}$. The Debye temperature is different for different solids being in the range of several hundred degrees. Obviously, the Debye temperature of the solid cannot exceed the Debye temperature of its optical branches: $\Theta<\Theta^O_j$. Write the energy of lattice vibrations as a function of temperature:

$$\langle E\rangle=\langle E^A\rangle+\langle E^O\rangle=\frac{3V(kT)^4}{2\pi^2c_0^3\hbar^3}\int\limits_0^{x_M}\frac{x^3\,dx}{e^x-1}+N\sum_{j=4}^{3s}\frac{k\Theta^O_j}{e^{\frac{\Theta^O_j}{T}}-1}. \tag{60.37}$$

Making use of (60.36) we may write the coefficient in front of the integral in (60.37):

$$\frac{3V(kT)^4}{2\pi^2c_0^3\hbar^3}=9kTN\left(\frac{T}{\Theta}\right)^3. \tag{60.38}$$

Subsequently $\langle E\rangle$ may be presented in the form

$$\langle E\rangle=NkT\left\{\frac{9}{\left(\frac{\Theta}{T}\right)^3}\int\limits_0^{x_M}\frac{x^3\,dx}{e^x-1}+\sum_{j=4}^{3s}\frac{\frac{\Theta^O_j}{T}}{e^{\frac{\Theta^O_j}{T}}-1}\right\}. \tag{60.39}$$

Since, by definition, $x=\frac{\hbar\omega}{kT}$, $x_{max}=\frac{\hbar\omega_{max}}{kT}$, but $\hbar\omega_{max}=k\Theta$, therefore, $x_{max}=\frac{\Theta}{T}$. Consequently,

$$\langle E\rangle=NkT\left\{\frac{9}{\left(\frac{\Theta}{T}\right)^3}\int\limits_0^{\frac{\Theta}{T}}\frac{x^3\,dx}{e^x-1}+\sum_{j=4}^{3s}\frac{\frac{\Theta^O_j}{T}}{e^{\frac{\Theta^O_j}{T}}-1}\right\}. \tag{60.40}$$

Consider the specific heat of the solid

$$C=\frac{\partial\langle E\rangle}{\partial T}. \tag{60.41}$$

The general expression for C is rather complicated. Consider high and low temperature ranges. When $T \gg \Theta$ and $T \gg \Theta_j^0$, the contribution of the optical vibrations is proportional to

$$\sum_{j=4}^{3s} \frac{\frac{\Theta_j^0}{T}}{e^{\frac{\Theta_j^0}{T}}-1} \cong \sum_{j=4}^{3s} 1 = (3s-3), \tag{60.42}$$

and the contribution of the acoustical vibrations is proportional to the value of the Debye function $D\left(\frac{\Theta}{T}\right)$:

$$D\left(\frac{\Theta}{T}\right) = \frac{3}{\left(\frac{\Theta}{T}\right)^3} \int_0^{\frac{\Theta}{T}} \frac{x^3\,dx}{e^x-1}. \tag{60.43}$$

For large T, when $\frac{\Theta}{T} \ll 1$ and $x \ll 1$, $e^x \cong 1+x$; $D\left(\frac{\Theta}{T}\right) \cong 1$ and, consequently,

$$\langle E\rangle \cong NkT[3+(3s-3)] = 3sNkT. \tag{60.44}$$

The energy per one degree of freedom is kT, and the specific heat of the lattice is independent of temperature:

$$C = 3Nsk. \tag{60.45}$$

The molar specific heat turns out to be 6 cal/mol which agrees with the Dulong-Petit law. The law is valid for metals up to $T < \frac{F}{k} = T_{deg}$ (T_{deg}—degeneracy temperature), and for semiconductors until the free electron concentration remains much lower than the concentration of the matrix atoms.

Consider another limiting case: $T \ll \Theta$ and, consequently. $T \ll \Theta_j^0$. This means that the upper limit of integration in the Debye integral may be put ∞, but since

$$\int_0^\infty \frac{x^3\,dx}{e^x-1} = \frac{\pi^4}{15}, \tag{60.46}$$

$$D\left(\frac{\Theta}{T}\right) = \frac{\pi^4}{5}\left(\frac{T}{\Theta}\right)^3 \tag{60.47}$$

and

$$\langle E\rangle = \frac{3\pi^4}{5}\frac{kT^4}{\Theta^3}. \tag{60.48}$$

The result is

$$C = \frac{12\pi^4 k}{5}\left(\frac{T}{\Theta}\right)^3. \tag{60.49}$$

The $C \sim T^3$ law works well in the temperature range up to

20-50 K: In the intermediate temperature range both the acoustical and the optical vibrations should be considered. In this range the correlation between the Debye theory and experiment is not very satisfactory. The reason is that short wave vibrations are excited for which the Debye interpolation is not valid. Indeed, Kellermann, who considered a linear atomic chain of the NaCl-type taking account of the dynamic matrix elements, obtained with the aid of numerical calculations a complex dependence ω (*K*) (or *K* (ω)) which he used to draw the curve presented in Fig. 84.

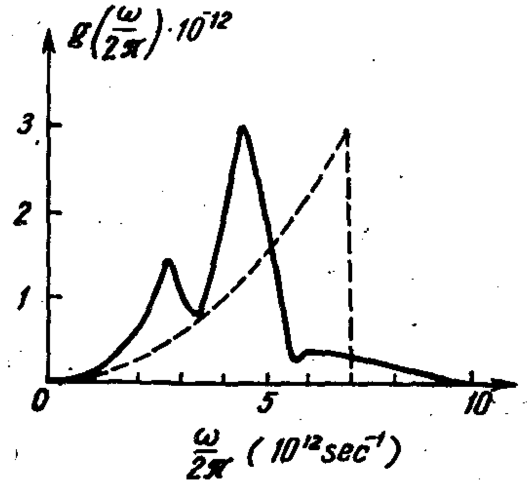


Fig. 84. The density of vibrations *g* (ω) for NaCl after Kellermann (solid line) and Debye (dashed line)

In conclusion of the section we would like to draw attention to two problems: to thermal lattice expansion and to thermal resistance. In the harmonic approximation of lattice vibrations *the normal co-ordinates are independent*. A definite set of phonons corresponds to each normal co-ordinate. They, too, should be independent, because *the phonons do not interact in the harmonic approximation*. The energy carried with the phonons travels through the lattice with the group velocity; consequently, *no resistance is presented to the heat flow*. The average displacement of the atoms vibrating harmonically is zero, therefore the interatomic distances remain unchanged. Real bodies are subject to thermal expansion. To explain this fact one should take into account the cubic terms of the expansion of the potential energy in atomic displacements. Should the tensor $B_{nj,\,ml,\,lk}$ be considered, the atomic vibrations would cease to be harmonic for the cubic term would make them anharmonic and interdependent. The oscillators corresponding to the normal coordinates would, too, become anharmonic. *The anharmonism of the oscillators results in the interacion of the phonons. This makes phonon-phonon scattering possible*, i.e. the superposition principle for the elastic waves is violated. Three phonons should take part in this interaction, and this could result both in the generation of a new phonon and in the annihilation of one of them (to be precise, two phonons merge into one). *The phonon-phonon scattering results in thermal resistance*. Since the anharmonism of the vibrations decreases with the decrease in the amplitude of vibrations, or in the temperature, the thermal resistance, too, should decrease with it. *The anharmonism results in average atomic displacements being non-zero and this, in turn, occasions changes in the dimensions of the bodies*, i.e. *thermal expansion of the solids takes place*.

Summary of Sec. 60

1. Statistical physics shows that the probability for some system to be in a state with the energy E_s is equal to

$$w_s = Z^{-1} e^{-\frac{E_s}{kT}}, \tag{60.1s}$$

where the statistical sum Z is determined from the condition of normalization for the probability:

$$\sum_s w_s = 1; \quad Z = \sum_s e^{-\frac{E_s}{kT}}. \tag{60.2s}$$

The average, or equilibrium, energy value $\langle E \rangle$ is

$$\langle E \rangle = \sum_s E_s w_s. \tag{60.3s}$$

2. For a harmonic oscillator the equilibrium energy value, according to (60.3s), is

$$\langle E_\alpha \rangle = \frac{\hbar\omega_\alpha}{2} + \frac{\hbar\omega_\alpha}{e^{\frac{\hbar\omega_\alpha}{kT}} - 1}. \tag{60.4s}$$

If the excitation of an oscillator is regarded as the generation of a phonon, the average number of phonons may be determined from the equation:

$$\langle n_\alpha \rangle = \frac{\langle E_\alpha \rangle}{\hbar\omega_\alpha} - \frac{1}{2} = \frac{1}{e^{\frac{\hbar\omega_\alpha}{kT}} - 1}. \tag{60.5s}$$

It follows from (60.5s) that phonons obey the Bose-Einstein statistics and, accordingly, have an integral spin (in $\hbar$ units).

3. Thus, phonons are quasiparticles corresponding to harmonic vibrations of the crystal lattice with the frequency ω_α and the wave vector $\mathbf{K}_\alpha$.

The energy of the phonon is

$$E_\alpha = \hbar\omega_\alpha, \tag{60.6s}$$

and the quasimomentum,

$$\mathbf{P}_{ph} = \hbar\mathbf{K}_\alpha. \tag{60.7s}$$

4. Characteristic temperature Θ of a solid, or its Debye temperature, is the term applied to a quantity determined by the condition:

$$\Theta = \frac{\hbar\omega_{max}^A}{k}. \tag{60.8s}$$

5. The Debye temperature for optical lattice vibrations is the term applied to the quantity Θ_j^O determined by the condition

$$\Theta_j^O = \frac{\hbar\omega_j^O}{k}. \tag{60.9s}$$

6. At temperature T the number of optical phonons is

$$n_j^O = \frac{1}{e^{\frac{\Theta_j^O}{T}} - 1}. \tag{60.10s}$$

7. The number of acoustical phonons depends on their frequency and on temperature in accordance with (60.5s). The rise in the number of long-wave phonons with the rise in temperature is more rapid than the rise in the number of short-wave phonons. At high temperatures $T \gg \Theta$ the energy carried with the phonons is independent of their frequency and is equal to kT per a degrée of freedom.

8. The specific heat of a solid at $T \ll \Theta$ is proportional to T^3; at $T \gg \Theta$ it is independent of temperature.

9. The anharmonicity of lattice vibrations results in thermal expansion of the solids and in thermal resistance, i.e. in phonon-phonon scattering.

61. SCATTERING BY THERMAL LATTICE VIBRATIONS. METHOD OF DEFORMATION POTENTIAL

The scattering of electrons and holes by lattice vibrations may be explained in terms of the corpuscular model with the aid of the phonon concept. Charge carriers colliding with phonons exchange energy and quasimomentum with it. Since the number of phonons depends on temperature, the charge scattering, too, should be temperature-dependent.

However, in order to apply the general theory of quantum transitions in Sec. 55, the perturbation generated by lattice vibrations should be found, which causes the transitjons of electrons and holes from state to state. To find this perturbation we intend to apply the method of the deformation potential.

A harmonic wave (ω, **K**) occasions local strains in the crystal. The displacement caused by the elastic wave at point **r** at the moment t is of the form

$$\mathbf{u}(\mathbf{r}, t) = \mathbf{u}_0 e^{-i\left[\omega t - (\mathbf{K}\mathbf{r}) + \frac{\pi}{2}\right]}, \tag{61.1}$$

where $\mathbf{u}_0$ is the vibration amplitude; the phase $\frac{\pi}{2}$ was introduced to simplify calculations.

The strain $\boldsymbol{u}$ results in relative variation of the volume Δ which, according to (51.17), is equal to div $\boldsymbol{u}$:

$$\Delta = \operatorname{div} \boldsymbol{u} = i\,(\mathbf{K}\boldsymbol{u}_0)\, e^{-i\left[\omega t - (\mathbf{Kr}) + \frac{\pi}{2}\right]} = (\mathbf{K}\boldsymbol{u}_0)\, e^{-i\,[\omega t - (\mathbf{Kr})]}. \quad (61.2)$$

As may be seen from (61.2) *the variation of volume in cubic crystals is caused only by longitudinal waves*. In anisotropic crystals waves of all types cause volume variations. Find the strain tensor which, according to (51.3), is equal to

$$\mathrm{u}_{mn} = \frac{1}{2}\left(\frac{\partial u_m}{\partial x_n} + \frac{\partial u_n}{\partial x_m}\right). \quad (61.3)$$

Since the m-component of the displacement $\boldsymbol{u}$ is determined by the m-component of $\boldsymbol{u}_0$, and since differentiation of (61.1) with respect to r_m results in the appearance of the factor K_m, we obtain

$$\mathrm{u}_{mn} = \frac{1}{2} e^{-i\,[\omega t - (\mathbf{Kr})]}\,(u_{0m}K_n + u_{0n}K_m). \quad (61.4)$$

Arrange the z-axis along the vector $\mathbf{K}$ with the result that $\mathbf{K} = (0, \theta, K)$. For a longitudinal wave $\boldsymbol{u}_0 = (0, 0, u_0)$; for transverse waves, $\boldsymbol{u}_0 = (u_0, 0, 0)$ or $\boldsymbol{u}_0 = (0, u_0, 0)$, depending on polarization. We obtain for the strain tensor

$$\mathrm{u} = u_0 K e^{-i\,[\omega t - (\mathbf{Kr})]} \times \begin{pmatrix} 0 & 0 & \theta \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (61.5)$$

$$\mathrm{u} = u_0 K e^{-i\,[\omega t - (\mathbf{Kr})]} \times \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad (61.6)$$

$$\mathrm{u} = u_0 K e^{-i\,[\omega t - (\mathbf{Kr})]} \times \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \quad (61.7)$$

Formula (61.5) *refers to the longitudinal wave*, and (61.6) and (61.7)—to the *transverse waves* polarized along x and y axes, respectively.

Since *the transverse waves represent a shear-type deformation, they do not result in volume variations* as may be seen from (61.6) and (61.7). Note, by the way, that since the longitudinal waves represent an alternation of compression and extension while the transverse waves represent a shear-type deformation, their velocities should be different, for Young's modulus and the shear modulus, generally, are not equal. *A localized strain caused by the* *wave results in a displacement of the bottom of the conduction band and of the top of the valence band.*

In compliance with (52.13) we write

$$E_c(\mathbf{u}) = E_c(0) + \sum_{ik} \Delta_{ik}^{(c)} \mathrm{u}_{ik} = E_c(0) + V_c(\mathbf{r}, t) \tag{61.8}$$

$$E_v(\mathbf{u}) = E_v(0) + \sum_{ik} \Delta_{ik}^{(v)} \mathrm{u}_{ik} = E_v(0) + V_v(\mathbf{r}, t), \tag{61.9}$$

where V_c and V_v are deformation potentials for the conduction and valence bands. Denote the element $\Delta_{33}^{(c)}$ by Δ_c and the element $\Delta_{13}^{(c)} = \Delta_{31}^{(c)} = \Delta_c'$. From considerations of symmetry write $\Delta_{23}^{(c)} = \Delta_{32}^{(c)} = = \Delta_c'$. For the sake of simplicity, we shall perform computations for the case of the longitudinal wave and of the conduction band:

$$V_c(\mathbf{r}, t) = \hat{\mathbf{W}}(\mathbf{r}, t) = u_0 K \Delta_c \cdot e^{-i[\omega t - (\mathbf{K}\mathbf{r})]}. \tag{61.10}$$

Calculate the matrix element of the perturbation operator (61.10) with the aid of the Bloch waves *in the effective mass approximation* and for normalization in a box of volume $G = L^3$:

$$\psi_{\mathbf{k}}(\mathbf{r}, t) = \psi_{\mathbf{k}}(\mathbf{r}) e^{-i\frac{E(\mathbf{k})t}{\hbar}} = \frac{1}{\sqrt{\upsilon}} e^{-i\left[\frac{Et}{\hbar} - (\mathbf{k}\mathbf{r})\right]}. \tag{61.11}$$

Write

$$W_{\mathbf{k}'\mathbf{k}} = V_{c\mathbf{k}'\mathbf{k}} = \frac{\Delta_c u_0 K}{G} e^{-\frac{i}{\hbar}(E - E' + \hbar\omega)t} \int_{(G)} e^{i(\mathbf{k} + \mathbf{K} - \mathbf{k},'\mathbf{r})} d\tau =$$
$$= \Delta_c u_0 K e^{-\frac{i}{\hbar}(E - E' + \hbar\omega)t} \delta_{\mathbf{k}+\mathbf{K},\mathbf{k}'}. \tag{61.12}$$

When calculating the integral in (61.12) we took into account the conditions of orthonormalization for the eigenfunctions $\psi_{\mathbf{k}}(\mathbf{r})$. The transition probability from the state $\mathbf{k}$ to the state $\mathbf{k}'$ is, according to (55.23), equal to

$$w(\mathbf{k}, \mathbf{k}') = \frac{2\pi}{\hbar} \Delta_c^2 u_0^2 K^2 \delta(E + \hbar\omega - E') \delta_{\mathbf{k}+\mathbf{K},\mathbf{k}'}. \tag{61.13}$$

The transition probability includes two δ-type quantities through the medium of which *the energy and quasimomentum conservation laws* are satisfied. In fact, *the probability will be non-zero only if*

$$E' = E + \hbar\omega; \quad \hbar\mathbf{k}' = \hbar\mathbf{k} + \hbar\mathbf{K}, \tag{61.14}$$

which means that *the energy and the quasimomentum of the electron are exchanged for the energy and quasimomentum of the phonon.* If one takes into account that the power of the exponent in (61.10) may be taken both with the plus and the minus sign, it turns out that (61.14) may be written in a more general form:

$$E' = E \pm \hbar\omega; \quad \hbar\mathbf{k}' = \hbar\mathbf{k} \pm \hbar\mathbf{K}, \tag{61.15}$$

i. e. the energy of the electron decreases or increases by the amount equal to the phonon energy; in other words, the electron either generates or absorbs a phonon.

Before proceeding with computations, note that the relations (61.15) may be interpreted in terms of wave concepts:

$$\frac{E'}{\hbar} = \frac{E}{\hbar} \pm \omega; \quad \mathbf{k}' = \mathbf{k} \pm \mathbf{K}, \tag{61.16}$$

i. e., the frequency $\frac{E}{\hbar}$ and the wave vector $\mathbf{k}$ of the electron wave change by an amount $\pm\omega$ and $\pm\mathbf{K}$ *as a result of the Doppler effect which takes place in the course of scattering of the electron by a moving object*, i. e. by the strain region.

The expression (61.13) is illustrated in Fig. 85 which shows also the constant-energy surfaces E and $E + \hbar\omega$. For a fixed phonon quasimomentum $\hbar\mathbf{K}$ two points A and B may be found on the energy surfaces E and E' such that the distance between them is $\mathbf{K}$ with the head of the vector $\mathbf{K}$ lying on the surface E' and the tail on the surface E. Such points may be several in number. The transition $A \to B$ corresponds to scattering with the absorption of a phonon. The points A' and B' correspond to electron scattering accompanied by the generation of a phonon. Denote the angle between $\mathbf{k}$ and $\mathbf{K}$ by α and express with its aid the relation between $\mathbf{k}$ and $\mathbf{K}$ taking into account (61.15):

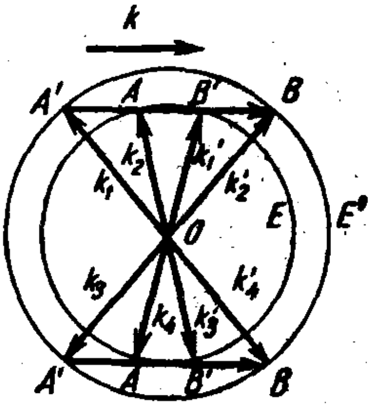


Fig. 85. Particle transitions obeying the energy and quasimomentum conservation laws

$$E' = E(\mathbf{k}') = \frac{\hbar^2 \mathbf{k}'^2}{2m^*} = \frac{\hbar^2}{2m^*}(\mathbf{k} \pm \mathbf{K})^2 = \frac{\hbar^2}{2m^*}(k^2 + K^2 \pm 2kK\cos\alpha) =$$
$$= E(\mathbf{k}) \pm \hbar\omega = \frac{\hbar^2 k^2}{2m^*} \pm \hbar\omega. \tag{61.17}$$

Cancelling $\frac{\hbar^2 k^2}{2m^*}$ in both members of the equation and substituting cK for ω *(this is justified for the long-wave part of the acoustical branch)* we obtain

$$K = \pm\frac{2m^* c}{\hbar} \mp 2k\cos\alpha. \tag{61.18}$$

Since $c \cong 10^5$ cm/s, the first term in (61.18) for $\frac{2m^*}{\hbar} \sim 1$ is equal to $\cong 10^{-5}$ cm^{-1}. The second term depends on the angle α; for small angles $\cos\alpha \cong 1$ and $K \cong 10^5 \mp 2k$. Express the quantity k in terms of temperature from the relation

$$\frac{\hbar^2 k^2}{2m^*} = \frac{3}{2} kT, \tag{61.19}$$

whence

$$k = \frac{1}{\hbar} \sqrt{3m^* kT}. \tag{61.20}$$

Setting $m^* \cong 10^{-27}$ g we obtain $k \cong 6\times 10^5 \sqrt{T}$ cm^{-1}. For an electron to have $k \cong 10^5$ cm^{-1} the body temperature should be $\cong 1$ K. In other words, the first term in (61.18) may for all practical purposes be neglected. Therefore we write

$$K \cong \mp 2k\cos\alpha. \tag{61.21}$$

The quasimomentum of the phonon may vary from 0 to $2k$, depending on α. For $T = 300$ K, $k \cong 10^7$ cm^{-1} and therefore K may assume the values from 0 to 2×10^7 cm^{-1}, which corresponds to phonon energies of from 0 to $\hbar cK \cong 10^{-27}\times 10^5 \times 10^7 = 10^{-15}$ erg $=$ $= 6\times 10^{-4}$ eV. This shows that the main part in scattering is played by the long-wave phonons whose energy is much smaller than that of the charge carriers. Therefore, we may assume the charge carrier scattering by the phonons to be elastic. In this case the phonon energy in the expression for $w(\mathbf{k}, \mathbf{k}')$ may be omitted.

Taking into account (61.21) transform the δ-function of energy as follows:

$$\delta(E \pm \hbar\omega - E') \cong \delta(E - E') = \delta\left(\frac{\hbar^2 k^2}{2m^*} - \frac{\hbar^2 k'}{2m^*}\right) = \frac{2m^*}{\hbar^2}\delta(k'^2 - k^2) =$$

$$= \frac{2m^*}{\hbar^2}\delta(K^2 \pm 2kK\cos\alpha) = \frac{m^*}{\hbar^2 kK}\delta\left(\frac{K}{2k} \pm \cos\alpha\right), \tag{61.22}$$

and obtain for the transition probability (61.13)

$$w(\mathbf{k}, \mathbf{k}') = \frac{2\pi m^* \Delta_c^2 u_0^2 K}{\hbar^3 k}\delta\left(\frac{K}{2k} \pm \cos\alpha\right). \tag{61.23}$$

To calculate relaxation time one should take into account the scattering by phonons of all possible K's. In other words, in the general case the displacements of the lattice atoms should be represented as a superposition of normal vibrations:

$$u(\mathbf{r}, t) = \sum_{\alpha} u_{0\alpha} e^{-i\left[\omega_\alpha t - (\mathbf{K}_\alpha \mathbf{r}) - \frac{\pi}{2}\right]}. \tag{61.24}$$

Write for the deformation potential

$$V_c = \Delta_c \sum_{\alpha} u_{0\alpha} K_{\alpha} e^{-i[\omega_{\alpha} t - (\mathbf{K}_{\alpha}\mathbf{r})]}. \tag{61.25}$$

Having found the deformation potential (61.25) we should now calculate the matrix element of the perturbation operator, and this will enable us to write the expression for the transition probability $w(\mathbf{k}, \mathbf{k}')$. Generally, the square of a sum is not equal to the sum of the squares, but it must be taken into account that the general expression for the probability will contain terms with products of δ-type quantities, which will drop out in the process of summation with the result that only the squares of the matrix elements of harmonic displacement components (61.24) will remain under the single summation sign.

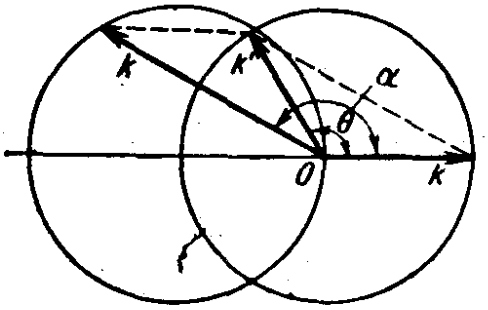


Fig. 86. The relation between the angles α and θ

Physically, this result is obvious. *Since the harmonic waves proportional to the normal co-ordinates are independent, the processes of carrier scattering by different waves are independent, too. Therefore the probabilities add up, and no interference effects take place.* Integrating (61.23) over all the k's with the weight $(1 - \cos\theta)$ we obtain $\tau^{-1}(\mathbf{k})$:

$$\frac{1}{\tau(\mathbf{k})} = \frac{1}{4\pi^3} \int\limits_{(V_{\mathbf{k}})} w(\mathbf{k}\mathbf{k}')[1 - \cos\theta]\, d\tau_{\mathbf{k}'}. \tag{61.26}$$

The integral (61.26) is conveniently calculated with the variable $\mathbf{K}$. Since $\mathbf{k}' = \mathbf{k} + \mathbf{K}$ it follows that $\mathbf{K} = \mathbf{k}' - \mathbf{k}$, and $d\tau_{\mathbf{k}'} = d\tau_{\mathbf{K}}$. We are considering elastic scattering, and for this reason the vectors $\mathbf{k}$ and $\mathbf{k}'$ lie on the same constant-energy surface (which is a sphere). It may be seen from Fig. 86 that for the vector $\mathbf{K}$ to assume all possible values for a fixed energy value, the vector $\mathbf{k}$ should assume values lying on the sphere of the radius k with the origin of co-ordinates at the point $(-\mathbf{k}')$. Denote the angle between $\mathbf{k}$ and $\mathbf{K}$ by α, and between $\mathbf{K}'$ and $\mathbf{k}$, by θ. The angle θ varies between 0 and π, the angle α—from $\frac{\pi}{2}$ to π, and $|\mathbf{K}|$—from 0 to $2|\mathbf{k}|$.

Find the relation between α and θ. Since $\cos\alpha = \frac{(\mathbf{kK})}{kK}$ and $\mathbf{K} = \mathbf{k}' - \mathbf{k}$, it follows that

$$\cos\alpha = \frac{(\mathbf{k}'\mathbf{k}) - \mathbf{k}^2}{kK} = \frac{k'k\cos\theta - k^2}{kK} = -\frac{k}{K}(1-\cos\theta), \qquad (61.27)$$

or

$$1-\cos\theta = -\frac{K}{k}\cos\alpha. \qquad (61.28)$$

Making use of (61.28) and (61.23) re-write (61.26) in the following form:

$$\frac{1}{\tau(\mathbf{k})} = -\frac{1}{4\pi^3}\frac{2\pi m^*\Delta_c^2}{\hbar^3 k}\int\limits_0^{2k} u_0^2(\mathbf{K})\,KK^2\,dK \int\limits_{\frac{\pi}{2}}^{\pi} \delta\left(\frac{K}{2k} \pm \cos\alpha\right) \times$$

$$\times \frac{K}{k}\cos\alpha\sin\alpha\,d\alpha \int\limits_0^{2\pi} d\varphi. \qquad (61.29)$$

Calculate the integral over α:

$$-\int\limits_{\frac{\pi}{2}}^{\pi} \delta\left(\frac{K}{2k} \pm \cos\alpha\right)\cos\alpha d\cos\alpha = \int\limits_0^{-1} \delta\left(\frac{K}{2k} \pm x\right)x\,dx = \frac{K}{2k} \qquad (61.30)$$

and

$$\frac{1}{\tau(\mathbf{k})} = \frac{m^*\Delta_c^2}{4\pi^2\hbar^3 k^3}\int\limits_0^{2k} u_0^2(\mathbf{K})\,K^5\,dK. \qquad (61.31)$$

In order to calculate the integral (61.31) the amplitudes of various vibrations should be known. To this end consider the *distribution of the oscillators among the vibration states*. At high temperatures all the oscillators are in high vibration states, their vibration energies coinciding with those of classical oscillators:

$$\frac{1}{2}M\omega^2 u_0^2 = kT, \qquad (61.32)$$

whence

$$u_0^2 = \frac{2kT}{M\omega^2} = \frac{2kT}{Mc^2K^2} \qquad (61.33)$$

and

$$\frac{1}{\tau(\mathbf{k})} = \frac{m^*\Delta_c^2}{4\pi^2\hbar^3 k^3}\int\limits_0^{2k} \frac{2kT\cdot K^5}{Mc^2K^2}\,dK = \frac{4m^*\Delta_c^2 kTk^4}{\pi^2 M\hbar^3 k^3 c^2} = \frac{4m^*\Delta_c^2 kT}{\pi^2 M\hbar^3 c^2}k, \quad (61.34)$$

i.e.

$$\tau(\mathbf{k})=\frac{\pi^2Mc^2\hbar^3}{4m^*\Delta_c^2kT}\cdot\frac{1}{k}=\frac{\pi^2Mc^2\hbar^4}{4\sqrt{2}(m^*)^{3/2}\Delta_c^2}\frac{E^{-1/2}}{kT}. \qquad (61.35)$$

The energy dependence of the relaxation time is of the form

$$\tau(\mathbf{k})=\tau(E)=\frac{\tau_0}{kT}E^p \text{ for } p=-\frac{1}{2}. \qquad (61.36)$$

The mean free path in atomic crystals in case of scattering by acoustical vibrations is independent of charge carries' energy $l \sim E^0$.

Should we define c by the relation $c^2=\frac{c_{ll}}{\rho}$ and set $M=\frac{\rho}{N}$, we would obtain

$$Mc^2=\frac{\rho}{N}\frac{c_{ll}}{\rho}=\frac{c_{ll}}{N}$$

and

$$\tau(E)=\frac{\pi^2\hbar^4c_{ll}}{2(2m^*)^{3/2}\Delta_c^2N}\frac{E^{-1/2}}{kT}. \qquad (61.37)$$

Hence, *the relaxation time is inversely proportional to the concentration N of the matrix atoms and directly proportional to the elasticity constant modulus* c_{ll}.

To describe the process of carrier scattering by optical vibrations use is made of the so-called *polarization potential*. It is introduced in the following way. The atoms of a cell vibrate in antiphase. The separation of charges results in an electric field being established (the substance is polarized), and this electric field travels through space in the form of a plane wave. The interaction of the charge carriers with this wave results in scattering. To find scattering probability the matrix element of the perturbation should be calculated. Should it be calculated with the aid of plane electron waves in the effective mass approximation, the result would be easily obtained that conservation laws should be observed in the process of scattering:

$$\hbar\mathbf{k}'=\hbar\mathbf{k}\pm\hbar\mathbf{K};\quad E'=E\pm\hbar\omega. \qquad (61.38)$$

The quantity $\hbar\omega$ should not be neglected since the energy of optical vibrations is quite large. Moreover, at sufficiently low temperatures the number of optical phonons will not be large, and kT cannot be substituted for u_0^2 *(u_0^2 is proportional to the number of optical phonons)*:

$$\langle n^O\rangle=\frac{1}{e^{\frac{\hbar\omega^O}{kT}}-1}. \qquad (61.39)$$

For this reason the temperature dependence of the relaxation time in case of scattering by optical vibrations should be of the form

$$\tau \sim \left[e^{\frac{\hbar\omega^O}{kT}} - 1\right]. \qquad (61.40)$$

At low temperatures this reduces to

$$\tau \sim e^{\frac{\hbar\omega^O}{kT}}. \qquad (61.41)$$

At high temperatures ($\hbar\omega^O \ll kT$) the number of phonons is proportional to kT just as for the acoustical phonons, but in this case τ is independent of the energy of charge carriers.

In real solids the mobility is determined by scattering both by acoustical and by optical phonons.

In conclusion of the paragraph we shall draw attention to another scattering mechanism which is operative in multivalley semiconductors. Since the lattice wave vector **K** changes from 0 to $\frac{2\pi}{\omega}$, *a carrier having absorbed a phonon may go over from one valley to another. Such processes are termed transfer processes.*

Summary of Sec. 61

1. An elastic wave produces strain in the crystal and thereby occasions displacement of the energy bands. The quantities V_c and V_v describing energy band displacements are termed deformation potentials. They may be expressed by the strain tensor **u** and the deformation potential coefficients tensor; for instance, for the conduction band

$$V_c(\mathbf{r}, t) = \sum_{ik} \Delta_{ik} \mathrm{u}_{ik}. \qquad (61.1s)$$

The strain tensor is expressed in terms of the displacement (61.2) by the formula (61.4).

2. The probability of charge carrier transition $w(\mathbf{k}, \mathbf{k}')$ caused by the deformation potential is of the form

$$w(\mathbf{k}, \mathbf{k}') = \frac{2\pi}{\hbar} \Delta_c^2 u_0^2 K^2 \delta(E \pm \hbar\omega - E')\, \delta_{\mathbf{k}+\mathbf{K},\, \mathbf{k}'}, \qquad (61.2s)$$

i.e. it is non-zero only in case the energy of an electron or a hole changes by the amount equal to the energy of a phonon. The simultaneous realization of the energy and quasimomentum conservation laws makes the following transformation of the expression (61.2s) possible:

$$w(\mathbf{k}\ \mathbf{k}')=\frac{2\pi}{\hbar^3}\frac{m^*\Delta_c^2 u_0^2K^2}{k}\delta\left(\frac{K}{2k}\pm\cos\alpha\right), \qquad (61.3s)$$

where α is the angle between **k** and **K**.

3. Since the processes of carrier scattering with normal co-ordinates remain independent of each other, performing the summation of (61.3) we obtain $\tau^{-1}(\mathbf{k})$. Integrating over α and K we obtain

$$\frac{1}{\tau(\mathbf{k})}=\frac{m^*\Delta_c^2}{4\pi^2\hbar^3k^3}\int\limits_0^{2k}u_0^2K^5\,dK. \qquad (61.4s)$$

To calculate the integral (61.4) one should know the distribution of the vibration amplitudes u_0 over frequencies or over K, i.e. one should know the numbers of phonons corresponding to definite frequencies. Generally, the number of phonons is determined by the Bose-Einstein function (60.10). Since at all finite temperatures a certain portion of the oscillators is excited to high quantum levels for which the number of phonons is $\frac{kT}{\hbar\omega}$, we may, by expressing the potential energy of vibrations in terms of phonon energy, obtain

$$\frac{1}{2}M\omega_\alpha^2u_0^2=\hbar\omega_\alpha\frac{kT}{\hbar\omega_\alpha}=kT. \qquad (61.5s)$$

From the standpoint of classical statistical physics the condition (61.5s) means that the energy per a degree of freedom equals kT. We obtain for $\tau(\mathbf{k})$

$$\tau(\mathbf{k})=\tau(E)=\frac{\pi^2\hbar^4Mc^2}{2(2m^*)^{3/2}\Delta_c^2}\frac{E^{-1/2}}{kT}=\frac{\pi^2\hbar^4c_{ll}}{2(2m^*)^{3/2}\Delta_c^2N}\frac{E^{-1/2}}{kT}. \qquad (61.6s)$$

4. Should the transport effective cross section be expressed in terms of the relaxation time, it would become evident that the transport effective cross section for scattering by thermal lattice vibrations is independent of the particles' energy. The scattering by thermal lattice vibrations is isotropic.

5. The scattering by optical lattice vibrations may be described with the aid of the polarization potential.

*62. TEMPERATURE DEPENDENCE OF CHARGE CARRIER MOBILITY

The results of the preceding sections enable the temperature dependence of charge carrier mobility to be considered:

$$\mu=\frac{e\langle\tau\rangle}{m^*}=\mu(T). \qquad (62.1)$$

The variations of temperature may result in the variations of the effective mass. The most elementary, but not the only cause of such a variation is the thermal expansion of the lattice. However, in the following we will take no account of the effective mass variations, but will confine ourselves to the consideration of the temperature dependence of the averaged relaxation time. According to (40.4),

$$\langle\tau\rangle=\langle E\tau\rangle=\frac{\int\limits_0^\infty x^{3/2}e^{-x}\tau(x)\,dx}{\int\limits_0^\infty x^{3/2}e^{-x}\,dx}, \tag{62.2}$$

where $x=\frac{E}{kT}$. The calculations will present no difficulties if the dependence of the relaxation time on the energy is a power-dependence, i.e. if

$$\tau(E)=\tau_0E^p=\tau_0(kT)^p x^p, \tag{62.3}$$

where p is the power and τ_0, a quantity independent of energy. Substituting (62.3) into (62.2) we obtain

$$\langle\tau\rangle=\tau_0(kT)^p\frac{\int\limits_0^\infty x^{3/2+p}e^{-x}\,dx}{\int\limits_0^\infty x^{3/2}e^{-x}\,dx}=\tau_0(kT)^p\frac{\Gamma\left(p+\frac{5}{2}\right)}{\Gamma\left(\frac{5}{2}\right)}. \tag{62.4}$$

The ratios of the Γ-functions for some values of p's are shown in Table 9 (p. 262). Substituting (62.4) into (62.1) we write

$$\mu=\langle\mu\rangle=\frac{e\tau_0k^p}{m^*}\frac{\Gamma\left(\frac{5}{2}+p\right)}{\Gamma\left(\frac{5}{2}\right)}T^p. \tag{62.4'}$$

For a temperature independent τ_0 the temperature dependence of mobility is analogous to the dependence of the relaxation time on the energy.

Consider now specific electron and hole scattering mechanisms.

1. Scattering by thermal lattice vibrations. Taking into account that for the case of scattering by thermal lattice vibrations $p=-\frac{1}{2}$ we write, in accordance with (62.4) and (61.37),

$$\mu(T)=\frac{\pi^2\hbar^4c_{ll}e\Gamma(2)}{2(2m^*)^{3/2}N\Delta_c^2m^*\Gamma\left(\frac{5}{2}\right)}\frac{1}{(kT)^{3/2}}=\frac{\pi^{3/2}\hbar^4ec_{ll}}{3\cdot2^{1/2}N\Delta_c^2k^{3/2}}\frac{1}{m^{*5/2}T^{3/2}}. \tag{62.5}$$

The expression (62.5) for mobility includes quantities of three types: (1) universal constants e, $\hbar$, k, and numerical factors; (2) the properties of the solid c_{ll}, N, Δ_c, m^*; (3) temperature. The mobility in a given solid depends only on temperature:

$$\mu = \mu_{0T} T^{-3/2} \tag{62.6}$$

i.e. *the mobility due to thermal lattice vibrations decreases with the rise in temperature.* The meaning of this expression is quite simple: as the temperature rises, the number of phonons increases in proportion to it, the mean velocity increasing in proportion to $\sqrt{T}$. As a result, the probability of electrons or holes colliding with the phonons rises in proportion to $T^{3/2}$. As the temperature falls, the carrier mobility increases. It should, however, be kept in mind that for $T \longrightarrow 0$ *the expression* (62.5) *becomes meaningless since the relation* (61.32) *is no longer applicable.*

It is noteworthy that *the value of mobility is inversely proportional to the* 5/2 *power of the effective mass.* This is substantiated by experiment: the carrier mobilities in which the effective mass is small are large as compared to the mobilities in solids in which the effective mass is large, electron mobility exceeding that of holes.

The factor determining the temperature dependence of mobility in the case of scattering by optical lattice vibrations is

$$\left[e^{\frac{\hbar\omega^0}{kT}} - 1\right]. \tag{62.7}$$

2. Scattering by impurity ions. According to (56.23) the dependence of relaxation time τ_I on the energy includes two terms: a power term and a logarithmic term:

$$\tau_I = \frac{\sqrt{2}\,\varepsilon^2 m^{*1/2} E^{3/2}}{\pi Z^2 e^4 N_I \ln\left[1 + \left(\frac{\varepsilon E}{N_I^{1/3} Z e^2}\right)^2\right]}. \tag{62.8}$$

When averaging this expression we take the logarithmic function out of the integral, ascribing to it the value which it assumes at the point of the integrand function maximum. To this end 3 kT should be substituted for E. After the logarithmic function has been taken out of the integral we may write

$$\langle\tau_I\rangle = \frac{8\sqrt{2}\,k^{3/2}\,\varepsilon^2 m^{*1/2}}{\pi^{3/2} Z^2 e^4 N_I} \frac{T^{3/2}}{\ln\left[1 + \left(\frac{3\varepsilon kT}{N_I^{1/3} Z e^2}\right)^2\right]}. \tag{62.9}$$

Expression (62.9) leads to an expression for mobility in case of impurity ion scattering:

$$\mu_I = \frac{8\sqrt{2}k^{3/2}}{\pi^{3/2}e^3}\,\frac{\varepsilon^2}{Z^2N_I m^{*1/2}}\,\frac{T^{3/2}}{\ln\left[1+\left(\frac{3\varepsilon kT}{N_I^{1/3}Ze^2}\right)^2\right]}. \tag{62.10}$$

The mobility of charge carriers scattered by impurity ions is determined by quantities of three kinds: (1) by universal constants and numerical factors; (2) by the properties of the solid N_I, ε, Z, m^*; (3) by temperature.

Investigate the expression for μ_I assuming all the quantities, with the exception of temperature, to be constant. At high temperatures we shall neglect the logarithmic dependence of μ_I on temperature and, therefore, we will be able to say that mobility increases with temperature as follows:

$$\mu_I \cong \mu_{0I}T^{3/2}. \tag{62.11}$$

As the temperature decreases, the mobility due to impurity ion scattering decreases too. Attention should, however, be drawn to the following fact: for

$$T^2 \ll \frac{N_I^{2/3}Z^2e^4}{9\varepsilon^2k^2} \tag{62.12}$$

the logarithmic dependence becomes important, and we should write

$$\mu_I \cong \frac{8\sqrt{2}e}{9\pi^{3/2}k^{1/2}}\,\frac{T^{-1/2}}{N_I^{1/3}m^{*1/2}}. \tag{62.13}$$

i.e. the mobility $\mu_I \sim \frac{1}{\sqrt{T}}$. This result follows directly from the fact that the relaxation time in the low temperature range is, according to (56.24), described by the dependence $\tau_I \sim E^{-1/2}$. Assess the temperature for which (62.12) holds. Putting $\varepsilon \cong 10$, $Z = 1$, we obtain

$$T^2 \ll 5\times 10^{-9}N_I^{2/3}. \tag{62.14}$$

As may be seen from (62.14), the temperature interval δT where the relation (62.13) is valid depends on the impurity ion concentration N_I: for $N_I \to 0$ $\delta T \to 0$. It must, however, be kept in mind that for $N_I \to 0$, $\mu_I \to \infty$ at any temperature. In conclusion it should be pointed out that N_I depends on temperature; for a semiconductor doped with impurity of one kind, at $T \to 0$ $N_I \to 0$. For a semiconductor doped with both the donor and acceptor impurities, owing to mutual impurity compensation, at $T \to 0$ N_I assumes a constant value equal to twice the minority impurity concentration, i.e. for $N_d \ll N_d$ $N_I = 2N_a$.

3. Scattering by neutral centres. According to (57.11) the relaxation time for the case of scattering by neutral centres is independent of the energy:

$$\tau=\frac{m^{*2}e^2}{20\varepsilon\hbar^3}\frac{1}{N_n}. \qquad (62.15)$$

It follows from here that the corresponding mobility should be of the form

$$\mu_n=\frac{e^3m^*}{20\varepsilon\hbar^3N_n}. \qquad (62.16)$$

It does not explicitly depend on temperature. As temperature rises, the electrically active impurity is ionized, and because of this at elevated temperatures μ_n is determined by the electrically inactive impurity, i.e. by an impurity which does not produce free electrons and holes.

For dislocation scattering $\tau\sim E^{-1/2}$, therefore

$$\mu_D\sim\frac{1}{\sqrt{T}}. \qquad (62.17)$$

4. The general case. After we have considered the working of the individual mechanisms in "pure" form turn now to the general case when *all the scattering mechanisms act simultaneously. Assuming all the mechanisms to be independent of each other we may assert that the total scattering probability is equal to the sum of probabilities of scattering by scattering centres of all types*. Hence, full relaxation time τ is of the form

$$\tau^{-1}=\frac{1}{4\pi^3}\int\limits_{(V_{\mathbf{k}})}\sum_i w_i(\mathbf{k},\mathbf{k}')(1-\cos\theta)\,d\tau_{\mathbf{k}'}=\sum_i\tau_i^{-1}, \qquad (62.18)$$

or

$$\tau=\left(\sum_i\tau_i^{-1}\right)^{-1}. \qquad (62.19)$$

Since the τ_i's depend on energy, τ, too, is a function of the energy which may be obtained if the explicit expressions for $\tau_i(E)$ are substituted into the formula. Having found $\tau(E)$ and averaged it over the energy we obtain the expression for mobility:

$$\mu_d=\langle\mu\rangle=\frac{e\langle\tau\rangle}{m^*}=\frac{e}{m^*}\left\langle\frac{1}{\sum\frac{1}{\tau_i(E)}}\right\rangle \qquad (62.20)$$

Analyze this expression. The addend terms in (62.20) and (62.19) play a different part under different conditions.

Suppose the partial relaxation times are equal. In this case

$$\tau_1=\tau_2=\ldots=\tau_r \text{ and } \tau_r=\frac{\tau_i}{r};\ \mu=\frac{\mu_i}{r}, \tag{62.21}$$

i.e. should r mechanisms with identical relaxation times be active, the full relaxation time and full mobility would be r times smaller than the partial. In actual fact, the probability of all the partial relaxation times being equal is meagre. Let the relaxation time of the lth mechanism be much smaller than that of the rest. Since in this case τ_l^{-1} exceeds all the other addends in (62.19), we may assert that $\tau\leqslant\tau_l$. In other words, *the full relaxation time is determined by the process the partial relaxation time of which is in the circumstances the smallest with the result that* $\langle\tau\rangle\leqslant\langle\tau_l\rangle$. Since τ_l's depend on energy, we should compare the role of the specific relaxation processes for specific carrier energies. However, the parts played by slow and fast charge carriers may be different in different processes. To obtain a correct result one should in this case assess the part played by the particles with the average energy of $\frac{3}{2}kT$, since the particles with the energy close to the average constitute the majority of the free charge carriers. Should $\langle E\rangle$ be substituted for E, we would obtain approximately

$$\tau_l(\langle E\rangle)=\tau_0(kT)^p\left(\frac{3}{2}\right)^p \tag{62.22}$$

instead of the true value

$$\langle\tau_l\rangle=\tau_0(kT)^p\frac{\Gamma\left(\frac{5}{2}+p\right)}{\Gamma\left(\frac{5}{2}\right)}. \tag{62.23}$$

Thus, to assess the part played by some scattering mechanism, the $\langle\tau_l\rangle$'s (or their equivalent $\langle\mu_l\rangle$'s) should be compared. This justifies the conclusion that at *high temperatures the main scattering mechanism is thermal lattice vibration scattering:* $\tau\sim E^p$ for $p=-\frac{1}{2}$. *Carrier mobility decreases with the rise in temperature as* $T^{-3/2}$. As temperature decreases, lattice vibrations become less important, the mobility increases, but simultaneously increases the contribution of impurity ion scattering with the result that *at low temperatures* $\tau\sim E^p$ *for* $p=\frac{3}{2}$ *and* $\mu\sim T^{3/2}$. In other words, if only the scattering by the impurity ions and thermal lattice vibrations is taken into account, the conclusion may be drawn that the mobility μ increases in proportion to $T^{3/2}$ with the rise in temperature, passes through a maximum, and then decreases as $T^{-3/2}$.

We would like to remind our readers that in the temperature range where the scattering by thermal lattice vibrations is predominant, the factor A of the Hall coefficient is equal to $\frac{3\pi}{8}$, which corresponds to $p=-\frac{1}{2}$; in the low temperature range $A = \frac{315\pi}{512}\left(p=\frac{3}{2}\right)$.

In the same way should the other kinetic coefficients be calculated.

In conclusion of the section consider some experimental data. Figure 87 shows the temperature dependence of electron and hole mobilities in gallium arsenide. The regions where the mobilities rise and fall and where they reach maximum values are clearly seen in the figure. Curves of the same type are observed for many solids. Figure 88 shows the temperature dependences of electron and hole mobilities in silicon in the range from —50°C to 250°C for different impurity concentrations. It may be seen that both in

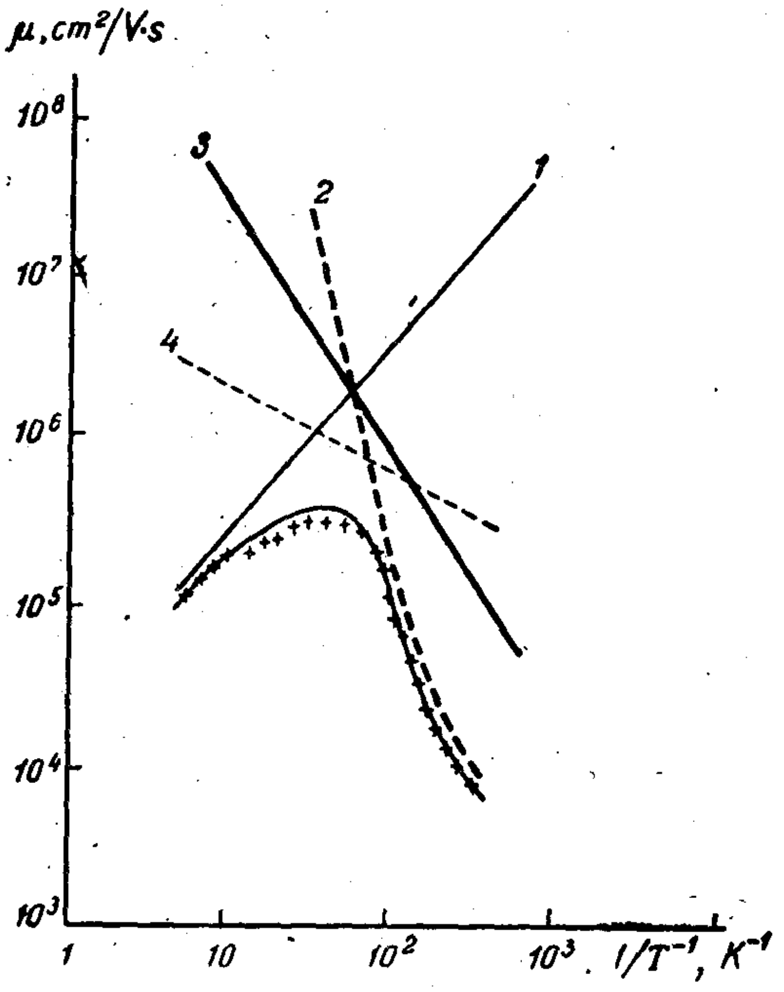


Fig. 87. The dependence of electron and hole mobilities on temperature in gallium arsenide for different scattering mechanisms:

1—impurity ions; 2—polar optical lattice vibrations; 3—deformation potential; 4—piezoelectric potential; solid line—total effect; xxx—experiment silicon and gallium arsenide the mobility of holes is less than that of electrons. This is true of all the semiconductors.

It may be seen from Figs. 87 and 88 that *the electron and hole mobilities decrease with rising impurity ion concentration.* Obviously, *in an intrinsic semiconductor of perfect crystal structure the mobility will be determined solely by thermal lattice vibrations. As the crystal becomes less perfect with the introduction of defects, the mobility must, of necessity, fall:* $\mu \leqslant \mu_T$. Figure 89 shows the room-temperature dependence of electron and hole mobilities in silicon on impurity concentration. It may be seen from Fig. 89 that for impurity concentrations of $N_I \leqslant 10^{14}$ cm^{-3} the mobility is practically independent of impurity ion concentration, but from $N_I \cong 10^{15}$ cm^{-3} onwards it begins to fall. In other words, at $T \cong 300°$C the ion impurity scattering becomes essential in silicon concentrations $N_I \geqslant 10^{15}$ cm^{-3}.

An important fact should be noted that *the mobility ratio* $b = \frac{\mu_n}{\mu_p}$ *decreases with the increase in impurity concentration.*

The dependence of electron and hole mobilities in A$^{\text{III}}$B$^{\text{V}}$ compounds on the impurity concentration is analogous. It is noteworthy that *electron mobility in a hole-type semiconductor is less than in an electron-type semiconductor.* The difference is the more pronounced the greater is the effective mass ratio $\frac{m_p^*}{m_n^*}$. The explanation is that, owing to a great difference in effective masses m_n^* and m_p^*, *the electrons are effectively scattered in the Coulomb field of the holes just as they are scattered in the field of the ions (scattering by free carriers).*

The above examples show that the carrier mobility may differ greatly from one semiconductor sample to another depending on

Table 20

Semiconductor	$\mu_n \sim T^p$	$\mu_p \sim T^p$	μ_n, cm^2 (V·s)$^{-1}$		μ_p, cm^2 (V·s)$^{-1}$		$b = \frac{\mu_n}{\mu_p}$	
	p	p	300 K	77 K	300 K	77 K	300 K	77 K
Germanium	—1.6	—2.3	3 800	37 100	1820	43 700	2.1	0.7
Silicon	—2.6	—2.3	1 300	45 500	500	11 600	2.6	3.9
InSb	—1.6	—2.1	78 000	1 200 000	750	10 000	100	120
InAs	—1.2	—2.3	33 000	82 000	460	690	70	120
InP	—2.0	—2.4	4 600	24 000	150	1 200	30	20
GaSb	—2.0	—0.9	4 000	6 000	1400	3 600	3	1.7
GaAs	—1.0	—2.1	8 500	21 000	420	4 200	20	5
GaP	—1.5	—1.5	110	500	75	420	1.5	1
AlSb	—	—1.8	200	—	420	3 700	—	—

its composition and crystal perfection. Nevertheless, we are justified in talking about *electron and hole mobility in a specific semiconductor. The numerical value of the mobility which is usually associated with the specific semiconductor refers to the purest and the*

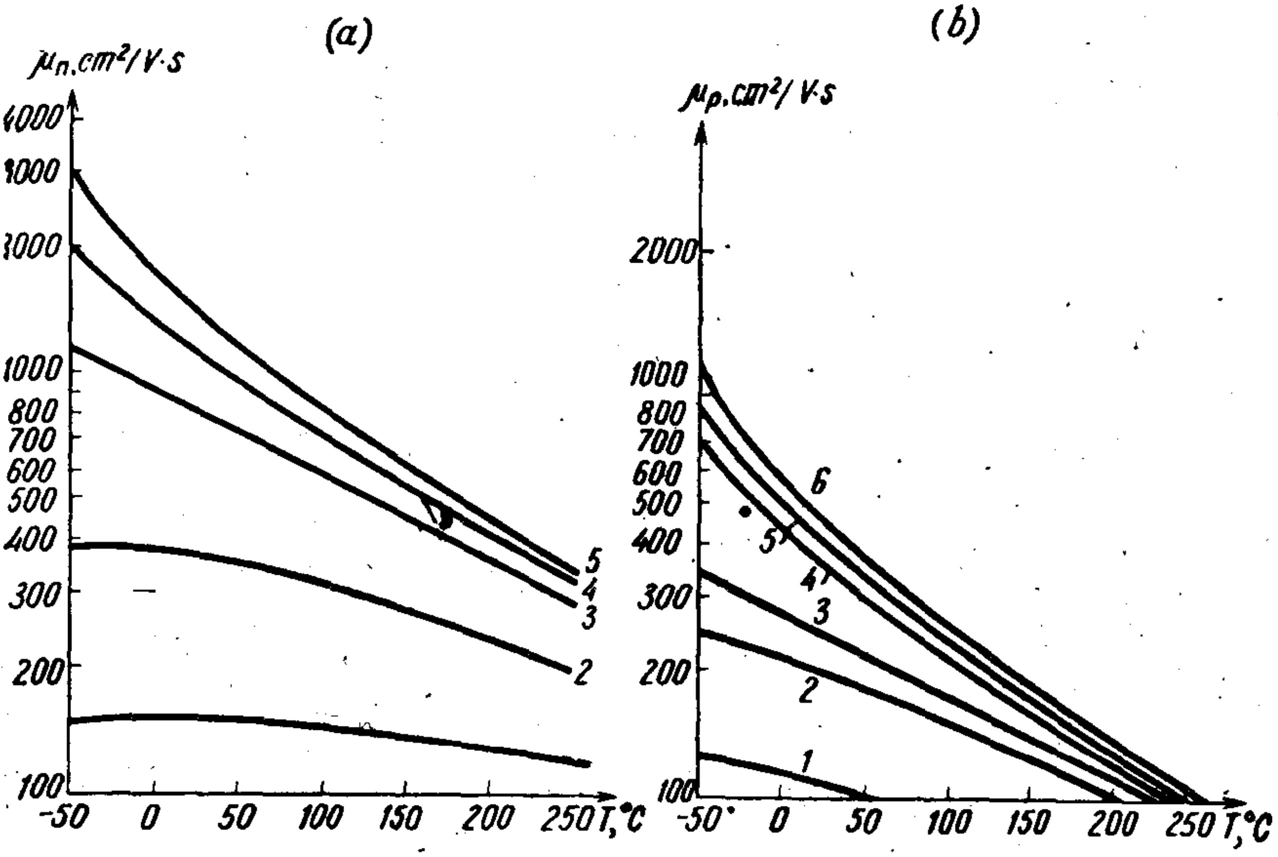


Fig. 88. The temperature dependence of the electron and hole mobilities in silicon with different impurity concentration N_I, cm^{-3}

(a) 1—5×10^{18}
2—1×10^{18}
3—1×10^{17}
4—1×10^{16}
5—1×10^{13}

(b) 1—5×10^{18}
2—1×10^{18}
3—5×10^{17}
4—1×10^{17}
5—1×10^{16}
6—1×10^{13}

most perfect single crystal. Table 20 shows the values of electron and hole mobilities and the mobility ratio *b* in a number of semiconducting materials at two temperatures: 300 K and 77 K.

Table 20 shows also the power *p* of the mobility temperature dependence $\mu \sim T^p$. It is worthy of note that only in some cases is the power *p* equal to the theoretical value of $\left(-\frac{3}{2}\right)$ corresponding to acoustical lattice vibration scattering. There is no satisfactory explanation for this discrepancy. Evidently, however, it reflects the fact that *there are different scattering mechanisms* leading to somewhat different temperature dependences for which the lattice vibrations are responsible. Indeed, computations show that two-phonon scattering should lead to a dependence $\mu \sim T^{-2}$. In intermetallic compounds of the $A^{III}B^{V}$ type an important part is played by the optical lattice vibrations (polarized vibrations). However, it may be said, that in spite of a large number of theoretical papers on the subject, the theory, as yet, is nowhere near

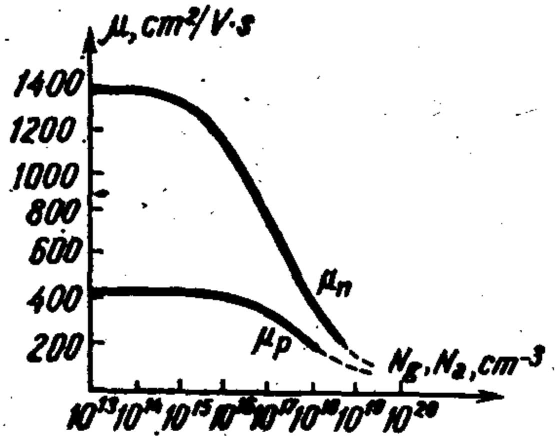


Fig. 89. The dependence of the electron and hole mobilities on impurity concentration in silicon at $T=300$ K

explaining the dependence of mobility in every specific solid on various factors and, primarily, on temperature.

In conclusion of the section let us stop to consider the temperature dependence of specific conductivity. The latter is determined

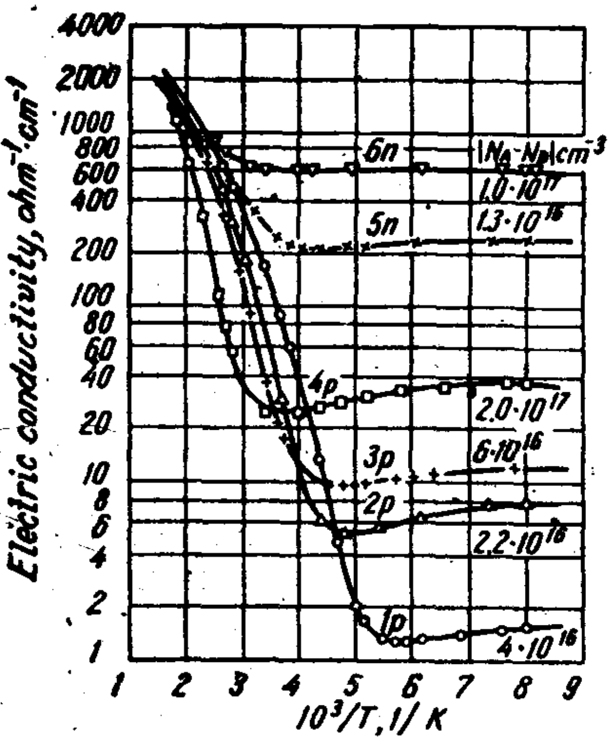


Fig. 90. The dependence of specific conductivity on inverse temperature in n- and p-silicon

by the dependence of carrier mobility and concentration on temperature:

$$\sigma(T)=\sum_{\alpha} e_{\alpha} n_{\alpha}(T)\, \mu_{\alpha}(T). \qquad (62.24)$$

In the impurity depletion range the concentration of majority carriers remains constant, and all the changes of conductivity are due only to the changes in mobility. The mobility may increase or decrease with temperature depending on impurity concentration, leading to the increase (or decrease) in the conductivity, respectively. In the temperature range where impurity ionization is low, or where the conductivity is intrinsic, the charge carrier concentration varies exponentially with temperature, and for this reason $\sigma(T)$ is determined by $n_\alpha(T)$. Figure 90 shows, by way of an example, the dependence of $\sigma(T^{-1})$ for different donor and acceptor concentrations in silicon.

Summary of Sec. 62

1. In non-degenerate semiconductors with one active scattering mechanism the averaged relaxation time $\langle\tau\rangle$ is determined by the relation:

$$\langle\tau\rangle=\tau_0\,(kT)^p\,\frac{\Gamma\left(p+\frac{5}{2}\right)}{\Gamma\left(\frac{5}{2}\right)}\qquad(62.1\text{s})$$

for $\tau(E)=\tau_0E^p$. When several mechanisms are active,

$$\tau^{-1}=\sum_i\tau_i^{-1}\qquad(62.2\text{s})$$

and

$$\langle\tau\rangle=\left\langle\left(\sum_i\tau_i^{-1}\right)^{-1}\right\rangle.\qquad(62.3\text{s})$$

In degenerate semiconductors $\langle\tau\rangle=\tau(F)$ and

$$\langle\tau\rangle=\left\langle\left\{\sum_i\tau_i^{-1}(F)\right\}^{-1}\right\rangle.\qquad(62.4\text{s})$$

2. For thermal lattice vibration scattering the mobility is proportional to $m^{*-5/2}$ and $T^{-3/2}$ when carriers are scattered by acoustical phonons; μ_T is proportional to $m^{*-3/2}$ and is determined by $\left(e^{\frac{\theta_j^O}{T}}-1\right)$ when carriers are scattered by optical vibrations.

3. For impurity ion scattering μ_I is proportional to $T^{3/2}$ and $m^{*-1/2}$, but at $T\to 0$ $\mu_I\sim T^{-1/2}$.

4. At elevated temperatures the main part in scattering is played by thermal lattice vibrations, and for this reason $A=\frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}=\frac{3\pi}{8}$. In the low temperature range the main part is played by impurity

ion scattering, and this results in $A=\frac{315\pi}{512}$; however, as T approaches zero, A decreases from $\frac{315\pi}{512}$ to $\frac{3\pi}{8}$. For the case of scattering by polarized vibrations A may vary from 1 at low temperatures to 1.1-1.3 at $T\sim\Theta_f^0$. For the case of scattering by neutral atoms μ is independent of temperature, and $A=1$. For electrons scattered by holes $\mu_n\sim T^{-3/2}$ and $A_n=\frac{315\pi}{512}$.

63. DEPENDENCE OF RELAXATION TIME ON EXTERNAL FIELDS. DEVIATIONS FROM OHM'S LAW

When solving the kinetic equation and analyzing the relationship between the relaxation time and the transition probability we made the point that in the simplest case the relaxation time

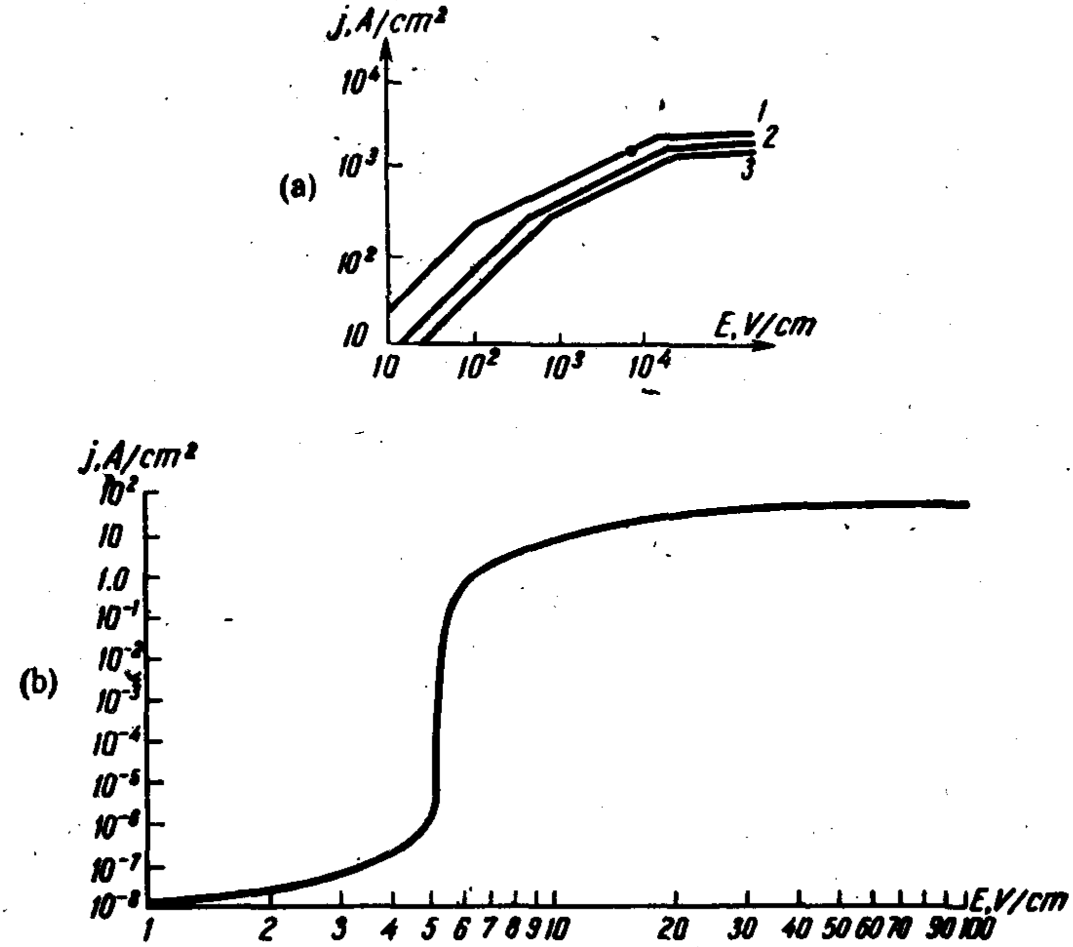


Fig. 91. The influence of strong electric fields on the conductivity of germanium at three different temperatures (*a*). Impact ionization in electron-type germanium at $T=4.2$. K (*b*).

should be independent of the fields. If, however, $\tau=\tau(\mathbf{E},\mathbf{B})$ the solution will be extremely complicated.

Figure 91 shows the dependence of the current density on the electric field intensity in electron-type germanium. The scale is double logarithmic. Three distinct regions are clearly visible in the graphs. In the first region, in fields up to 10^3 V/cm (at 77 K) the graphs are linear with an inclination of 45°, i.e. the slope is equal to unity; in the second region the slope is 1/2, and in the third it is zero. This means that in the first region σ is independent of the field, in the second $\sigma \sim E^{-1/2}$, and in the third $\sigma \sim E^{-1}$, so that in the respective fields $j \sim E$; $j \sim E^{1/2}$, and $j \sim E^0$. But since $\sigma = en\mu = \frac{e^2 n \langle\tau\rangle}{m^*}$, the deviations from Ohm's law may result from two causes; variations of $\langle\tau\rangle$ or of n. It may be easily understood how the field changes $\langle\tau\rangle$.

In case of isotropic scattering the relaxation time is practically equal to mean free time since the directional velocity is almost totally lost after one collision. The directional velocity depends solely on the field. The energy transmitted by the carrier in one collision is only a small fraction of its total energy: $\langle\Delta E\rangle = \frac{2m^*}{M} E$. However, since the particle system is in a stationary state, the electrons and holes give up only their excess energy. This means that during free transit time the particles receive from the field the same amount of energy $\frac{2m^*}{M} E$; consequently, the electric field little affects the velocity of the electron or of the hole. With the increase in the field full velocity and energy increase too. Since in case of thermal lattice vibration scattering the mean free path is independent of energy, the increase in velocity results in the decrease in the mean free time and, consequently, in mobility.

Find the variation of the thermal velocity v_T occasioned by the field. The field accelerates the particle so that energy received by it per unit time is equal to the work performed by the force $e\mathbf{E}$:

$$\frac{dE}{dt} = (e\mathbf{E}\mathbf{v}_d) = e\mathbf{E}\mu_d\mathbf{E} = \frac{e^2\tau}{m^*}\mathbf{E}^2. \tag{63.1}$$

Because of scattering the particle loses an amount of energy $\frac{2m^*}{M}E$ during the time τ so that, on the average, the energy $\frac{2m^*}{M}\frac{E}{\tau}$ is lost per unit time. Since the process is stationary, the average charge carrier energy remains constant in time:

$$\frac{dE}{dt} = \frac{2m^*}{M}\frac{E}{\tau}, \tag{63.2}$$

or

$$\frac{e^2\tau}{m^*}\mathbf{E}^2 = \frac{2m^*}{M}\frac{m^*v^2}{2\tau}. \tag{63.3}$$

Substituting $\frac{l}{v}$ for τ we obtain

$$e^2 l^2 \mathrm{E}^2 = \frac{m^{*2}}{M} v^4. \tag{63.4}$$

Let $l = l_0 v^{2r}$; in this case

$$\frac{e^2 M l_0^2 \mathrm{E}^2}{m^{*2}} = v^{(4-4r)} \tag{63.5}$$

and

$$v \sim \mathrm{E}^{\frac{1}{2(1-r)}}. \tag{63.6}$$

For the case of thermal lattice vibration scattering the mean free path is independent of energy; therefore, $r = 0$ and

$$v \sim \mathrm{E}^{1/2}. \tag{63.7}$$

An increase in the charge carriers' velocity, according to (63.6), results in the **decrease** in τ:

$$\tau \cong \frac{1}{v} \sim \frac{1}{\mathrm{E}^{1/2}}$$

whence

$$\mu = \frac{\mu_0}{\sqrt{\mathrm{E}}} \tag{63.8}$$

and

$$\sigma = \frac{\sigma_0}{\sqrt{\mathrm{E}}}; \quad j = \sigma_0 \sqrt{\mathrm{E}} \text{ for } \mathrm{E} > \mathrm{E}_{cr}. \tag{63.9}$$

The critical field E_{cr} is defined as a field for which the additional velocity imparted by the field to the particle becomes comparable to its thermal velocity. Clearly, the critical field E_{cr} should decrease with the decrease in temperature (Fig. 91). It is also obvious that the greater is μ_d, the less E_{cr}.

For $r \neq 0$ the dependence of v on the field is more intricate. For $r = 2$, as is the case for impurity ion scattering, we have, according to (63.6),

$$v \sim \mathrm{E}^{1/2}; \quad \tau \sim \mathrm{E}^{5/2}.$$

W. Shockley obtained the following expression for the dependence of mobility on the electric field intensity for the case of carrier-phonon interaction:

$$\mu(\mathrm{E}) = \mu_0 \frac{\sqrt{2}}{\sqrt{1 + \sqrt{1 + \frac{3\pi}{8}\left(\frac{\mu_0 \mathrm{E}}{c}\right)^2}}}, \tag{63.10}$$

where μ_0 is the mobility in weak fields, c—the speed of sound. For silicon the speed of sound is $\cong 7\times10^5$ cm/s. The field which imparts to electrons the speed equal to that of sound is $\cong 5.5$ kV/cm. For $\mathbf{E} \gg \frac{c}{\mu_0}$ we obtain from (63.10) the already familiar result $\mu \cong \mu_0 \frac{4}{\sqrt{3\pi}}\left(\frac{c}{\mu_0 \mathbf{E}}\right)^{1/2}$, i.e. $\mu \sim \mathbf{E}^{-1/2}$ which disagrees with Ohm's law.

Figure 92 shows an experimental dependence of drift velocity on electric field intensity.

Should the field be further increased beyond the values at which the current saturates, a value E'_{cr} would be reached at which a rapid increase in the current takes place.

This increase may be described by an empirical field dependence of the current:

$$\sigma = \sigma_0 e^{\gamma(\mathbf{E}-\mathbf{E}'_{cr})}. \tag{63.11}$$

This rapid increase in conductivity in an electric field is termed Poolè effect. The exponential rise in conductivity is usually the result of the increase in carrier concentration. There are several grounds for such an increase.

One of the mechanisms is termed *impact ionization*. The electron (or hole; receives during its mean free time an energy sufficient to ionize an impurity atom or even a matrix atom. The impact of a "hot" particle on an impurity or a matrix atom results in the generation of additional carriers which, in their turn, are accelerated by the field and generate upon impact on the atoms new free charge carriers. If one takes into account that the increase in carrier concentration is accompanied by the increase in their recombination one can easily envisage the *possibility of such a stationary state when the carrier concentration induced by the field and, consequently, the conductivity remain at a constant high level. Should recombination fail to compensate for the generation, the concentration would rise in an avalanche, and an irreversible break-down would take place in the semiconductor.*

Figure 91*b* shows the electric field dependence of the current density in electron-type germanium at $T = 4.2$ K. At $E \cong 5$ V/cm a jump in the current density takes place due to impact ionization of impurity atoms. However, as the field is further increased, the current remains unchanged. The phenomenon of impact ionization may be the result of internal fields produced by localized inhomogeneities of the crystal, or by *p-n* junction fields. The fields at which impact ionization takes place are the smaller the lower is the temperature, the smaller is the activation energy and the greater is the carrier mobility.

The second mechanism responsible for enhanced conductivity is

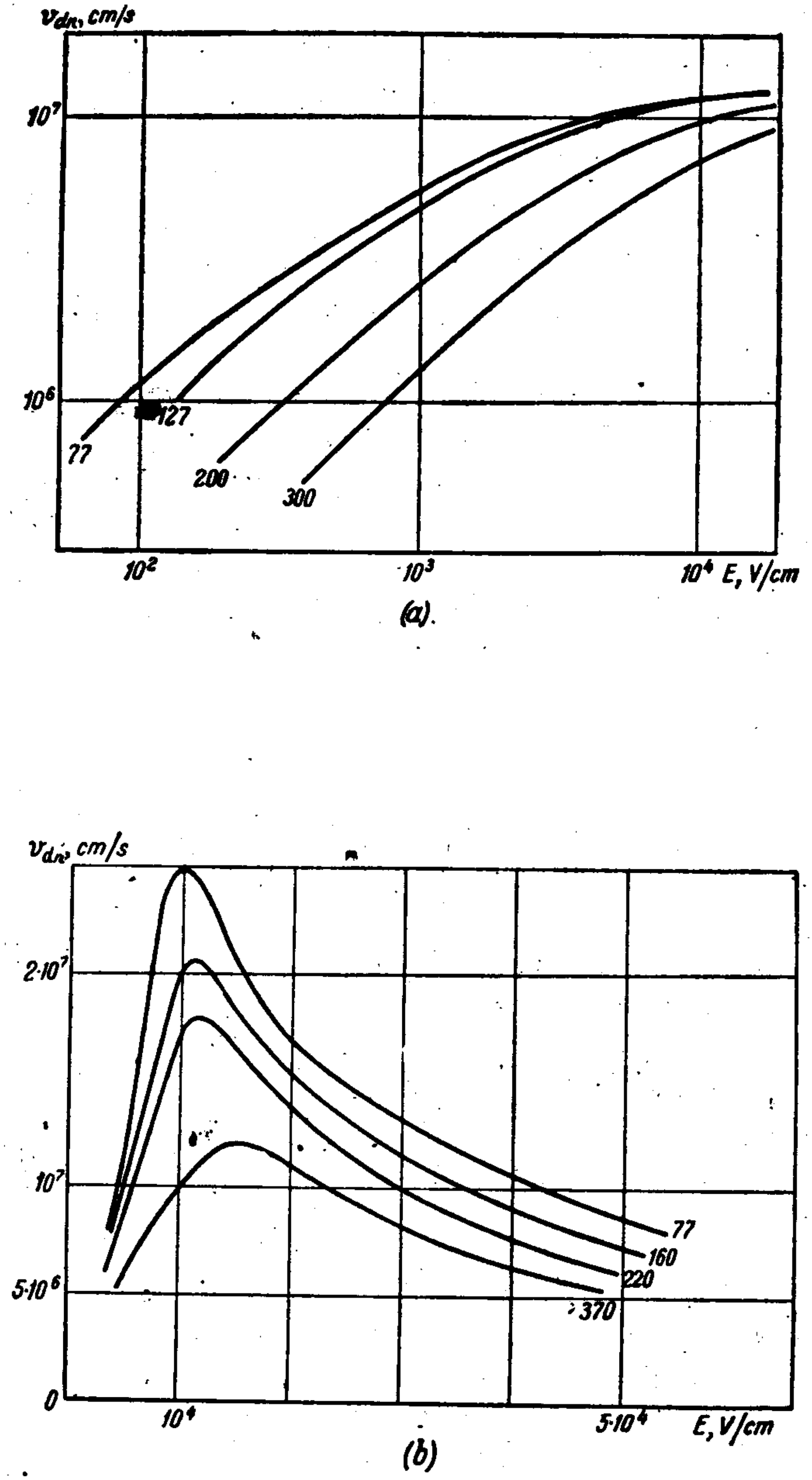


Fig. 92. The dependence of the drift velocity on electric field intensity in silicon (*a*) and cadmium telluride (*b*) at different temperatures, K

the Zener effect discussed in Sec. 19. The probability of tunnelling depends on the height ΔE_0 and the width a of the potential barrier. The barrier width, in turn, depends on the electric field intensity:

$$ae\mathrm{E} = \Delta E_0; \quad a = \frac{\Delta E_0}{e\mathrm{E}}. \tag{63.12}$$

It is established in quantum mechanics that the transition probability for a triangular barrier is of the form

$$D = D_0 e^{\frac{-2\sqrt{m^*}(\Delta E_0)^{3/2}}{2e\hbar \mathrm{E}}} \tag{63.13}$$

The tunnelling probabilities from the valence into the conduction band and vice versa are the same. However, since electron concentration is higher in the valence band, *the electron flow will be from the valence into the conduction band.* The Zener effect is analogous to cold emission. It is observed in fields of E>(10⁴-10⁵)V/cm. Thus, *the Zener effect results in the increase in free carrier concentration.*

The electrostatic ionization which is manifested in the Zener effect may also lead to other effects. One of them is *the Stark effect* resulting in *the widening of the energy levels of the atoms* constituting the crystal. This widening is accompanied by *the widening of the energy bands and the narrowing of the forbidden bands.* But the narrowing of the forbidden band results in an increase in electron and hole concentrations. The increase in carrier concentration occasioned by strong fields may, however, according to Frenkel, be explained *without resorting to the Stark effect.* Figure 93 shows the potential energy of the electron in an atom in the absence (dashed line) and in the presence of an electric field (solid line). The external field

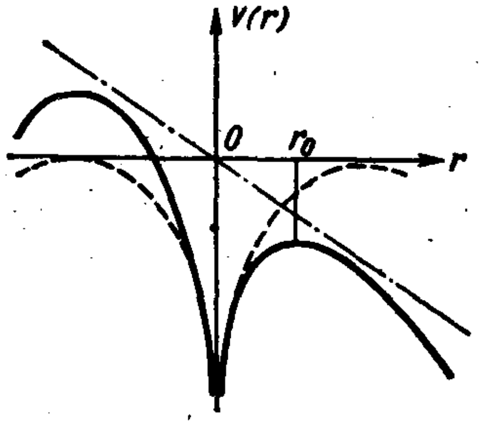


Fig. 93. The lowering of the potential barrier under the influence of an electric field

lowers the barrier in the direction opposite to the direction of the field. The potential field maximum is at the point r_0 where the force of attraction of the nearest atom is balanced by the external force:

$$\frac{e^2}{\varepsilon r_0^2}=e\mathrm{E}, \tag{63.14}$$

whence

$$r_0=\sqrt{\frac{e}{\varepsilon\mathrm{E}}}. \tag{63.15}$$

The decrease in the height of the potential barrier separating two neighbouring lattice sites may be estimated as

$$\delta E_1=2e\mathrm{E}r_0=2e\sqrt{\frac{e\mathrm{E}}{\varepsilon}}. \tag{63.16}$$

But the lowering of the potential barrier increases, according to Boltzmann's statistics, the probability of thermal excitation by the amount

$$e^{\frac{\delta E_1}{kT}}=e^{\left(\frac{2}{kT}\sqrt{\frac{e^3}{\varepsilon}}\sqrt{\mathrm{E}}\right)}=e^{\beta'\sqrt{E}}. \tag{63.17}$$

This effect is noticeable at E > 10^5-10^6 V/cm.

High-frequency current oscillations upon application of strong electric fields are observed in some solids. In gallium arsenide the frequency of the current oscillations is $(1\text{-}6)\times10^9\ \mathrm{s}^{-1}$, and the amplitude may exceed 1 A. Figure 94 shows current oscillogram of a sample of n-GaAs 0.025 mm long, to which a voltage pulse of 16 V and of 10^{-8} s duration has been applied. The upper part of the figure presents an exploded view of the oscillogram. *High-frequency current oscillations in semiconductors accompanying the application of constant voltages are termed Gunn effect*. This effect is observed in gallium arsenide, gallium phosphide and in some other solids.

One of the possible explanations of the Gunn effect is as follows. As the electric field is applied, the electrons go over to higher energy states, and the temperature of the electron gas rises. Suppose that some energy minima of the conduction band lie above the absolute minimum and that their effective masses are appreciably greater than that of the absolute minimum as is the case with gallium arsenide shown in Fig. 42. Electrons interacting with the phonons may go over to higher valleys. Since the density of states in the upper valley is greater than that in the lower valley, the electrons will be accumulated in the former. But the mobility of the electrons in the upper valley is substantially less than that of the electrons in the lower valley, therefore electron drift velocity will fall together with their contribution to the current, and the current, itself will fall, too. The states in the upper valley are unstable. The electrons interact with the phonons and go over to the lower valley increasing the current. Periodic oscillations are usually observed in thin samples, this being attributed to the mechanism of resistance increase in semiconductors. The greater the field intensity, the more intense is the process of electron transfer to the upper valley; but the more electrons go over to the upper valley, the greater will be the resistance of the particular region of the semiconductor, and the greater will be the potential drop in this region, leading to the decrease of electric field intensity in the nearby regions. Experiments show this increased resistance region to start at the "cathode" and to move towards the "anode". The Gunn effect is observed at field intensities at which the drift velocity becomes comparable to the thermal velocity, i.e. at $v_d \cong 10^7$ cm/s at room temperature.

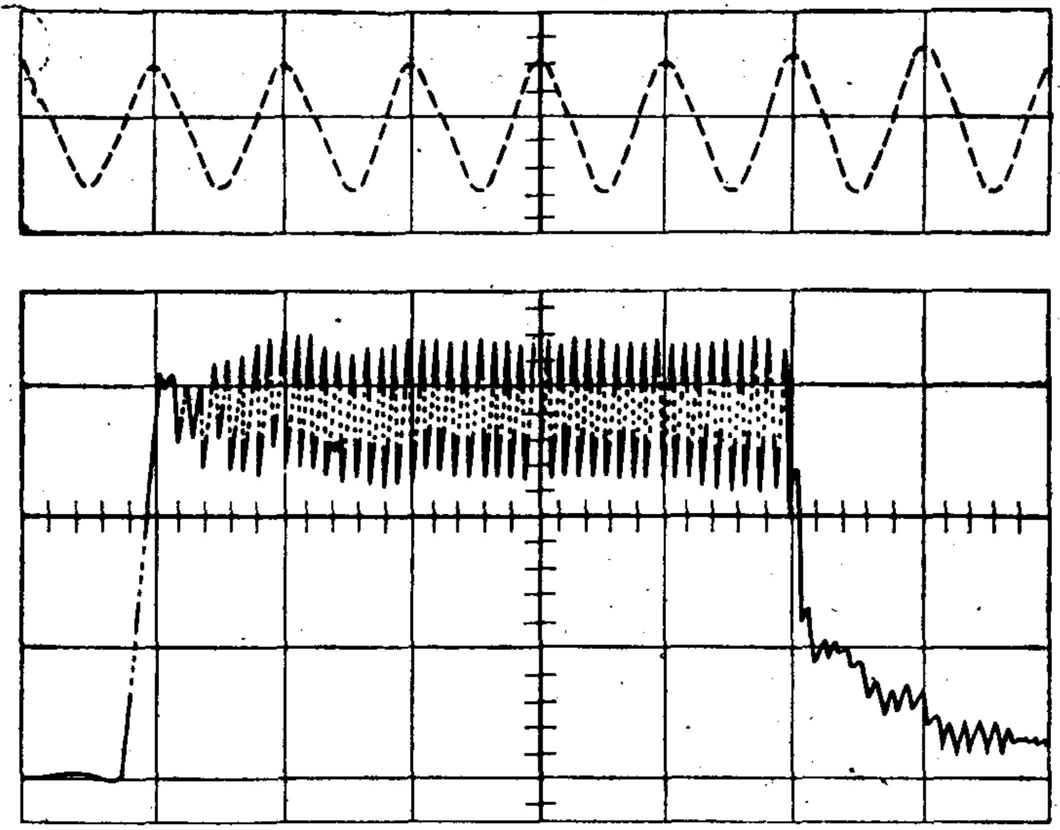


Fig. 94. Current oscillations in a gallium arsenide sample when a voltage pulse of 10^{-8} s duration is applied to it

The Gunn effect may be used to study the band pattern of semiconductors and to construct high-power oscillators for frequency bands of the order of 10^9 Hz.

Summary of Sec. 63

1. The external electric field changes the charge carrier energy with the resulting change in their mobility. The mobility may either increase or decrease, depending upon the scattering mechanism.

2. Strong electric fields occasion free carrier concentration variations through impact ionization, Zener effect (or internal cold emission), the effect of thermal excitation due to the lowering of the potential barrier, and Stark effect.

3. The transition of hot electrons to higher valleys, where their effective mass is greater than that in the lower valley, results in a decrease in electron drift velocity and, therefore, in the current, as well. The transition of the electrons to the lower valley leads to current increase. Periodic current oscillations in a semiconductor to which a constant voltage has been applied are termed Gunn effect.

Chapter VI

CHARGE CARRIER RECOMBINATION

64. CONTINUITY EQUATION. LIFETIME

In Chapter III we discussed in detail the dependence of charge carrier concentration on temperature, impurity type, and concentration. This was done in the assumption that the semiconductor was in a state of thermodynamic equilibrium. Charge carrier concentration corresponding to the state of equilibrium we shall term *equilibrium concentration*. In Chapter IV, when considering kinetics phenomena, it was assumed that external fields disturb the state of thermodynamic equilibrium. *The semiconductor carrying current is in a non-equilibrium state*. However, *the charge carrier concentration in a uniform semiconductor in isothermic condition remains equilibrium provided the fields are not too high*. But since the electric field disturbs the equilibrium, the charge carriers become non-equilibrium carriers, for their distribution among states is now described by a non-equilibrium distribution function. Indeed, according to (29.1s-2s) the expression for carrier concentration assumes the form

$$n(\mathbf{r})=\frac{1}{4\pi^3}\int\limits_{(V_\mathbf{k})} f(\mathbf{r},\mathbf{k}')\,d\tau_{\mathbf{k}'}=\frac{1}{4\pi^3}\int\limits_{(V_\mathbf{k})}[f_0(\mathbf{r},\mathbf{k})+f^{(1)}(\mathbf{r},\mathbf{k})]\,d\tau_\mathbf{k}=$$

$$=n_0(\mathbf{r})+n_1(\mathbf{r}),\qquad(64.1)$$

where n_0 denotes *the equilibrium electron (or hole) concentration*. Calculate n_1:

$$n_1(\mathbf{r})=\frac{1}{4\pi^3}\int\limits_{(V_\mathbf{k})} f^{(1)}(\mathbf{r},\mathbf{k})\,d\tau_\mathbf{k}=-\frac{1}{4\pi^3}\int\limits_{(V_\mathbf{k})}\frac{\partial f_0}{\partial E}(\chi\mathbf{v})\,d\tau_\mathbf{k}.\qquad(64.2)$$

It may easily be seen that $n_1(\mathbf{r})=0$. Indeed, χ and $\frac{\partial f_0}{\partial E}$ are even functions of $\mathbf{k}$, and $\mathbf{v}$ is an odd function; therefore, the integrand function in (64.2) is odd, and an integral of an odd function over symmetrical limits is zero. Hence,

$$n(\mathbf{r})=n_0(\mathbf{r}),\qquad(64.3)$$

i. e. *the concentration remains equilibrium*. It may be easily demonstrated that (64.3) follows from the condition (36.6): $\frac{df}{dt}=0$. The same condition leads to *the continuity equation* well known in hydrodynamics and in the theory of electricity:

$$\frac{d\rho}{dt}=\frac{\partial\rho}{\partial t}+\operatorname{div}\rho\mathbf{v}=\frac{\partial\rho}{\partial t}+\operatorname{div}\mathbf{j}=0, \tag{64.4}$$

where ρ is the mass (or charge) density, and $\mathbf{v}$ its velocity. The equation (64.4) means, for example, *that volume charge density may change only as a result of the current divergence*. The equation is quite general. However, if it is written only for carriers of *one* type its form will change appreciably *since the numbers of carriers of each type*, as was shown in Sec. 63, *may vary* with the total charge remaining constant. To describe such situations the Boltzmann equation, introduced in Chapter IV, should be generalized. However, we shall instead generalize the continuity equation. To begin with, introduce some definitions which will help us to understand numerous phenomena. *Electron or hole concentrations* $n(\mathbf{r}, t)$ *not equal to their equilibrium values* $n_0(\mathbf{r})$ *are termed nonequilibrium concentrations*. The quantity $\delta n(\mathbf{r}, t)$ equal to

$$\delta n(\mathbf{r}, t)=n(\mathbf{r}, t)-n_0(\mathbf{r}), \tag{64.5}$$

is termed *excess charge carrier concentration*.

The appearance of a pair of free carriers is termed pair generation; the reverse process, resulting in the annihilation of a pair of free charge carriers (of an electron and a hole), is termed recombination. The generation and recombination may be the result of internal and external causes. Denote *the concentration variation rates due to generation* resulting from internal causes by g_I, and from external causes, by g_E; *the concentration variation rates due to recombination* resulting from internal causes, by r_I and from external causes, by r_E. Write the continuity equation for electrons and holes taking into account their current, generation and recombination:

$$\frac{\partial n}{\partial t}=-\operatorname{div}\frac{\mathbf{j}_n}{e_n}+g_I^{(n)}+g_E^{(n)}-r_I^{(n)}-r_E^{(n)}, \tag{64.6}$$

$$\frac{\partial p}{\partial t}=-\operatorname{div}\frac{\mathbf{j}_p}{e_p}+g_I^{(p)}+g_E^{(p)}-r_I^{(p)}-r_E^{(p)}. \tag{64.7}$$

If one sets the initial and boundary conditions he may, knowing the dependences of $\mathbf{j}$, r and g on co-ordinates and on time, obtain, in principle, the solutions of the equations (64.6) and (64.7) for $n(\mathbf{r}, t)$ and $p(\mathbf{r}, t)$. However, a general solution of the continuity equation presents unsurmountable difficulties.

Consider some of the simpler cases. In a semiconductor in adiabatic conditions $g_E = r_E = 0$. In the absence of current ($j = 0$) the continuity equation (64.6) assumes the form

$$\frac{\partial n}{\partial t} = g_l - r_l. \tag{64.8}$$

For stationary conditions $\frac{\partial n}{\partial t} = 0$, and

$$g_l = r_l, \tag{64.9}$$

i.e. *carrier recombination and generation compensate each other maintaining the stationary state.* Since in this case the carriers in the process of their generation and recombination exchange energy with the thermal lattice vibrations, the values of g_l and r_l, which we denote by g_0 and r_0, determine the rates of carrier concentration variations due to *thermal generation and recombination.* The values of n and p, in this case denoted by n_0 and p_0, represent the equilibrium concentrations sometimes termed *dark current carrier concentrations.* The values of n_0 and p_0 are determined solely by the properties of the semiconductor. For example, for a non-degenerate semiconductor

$$n_0 = N_c e^{-\frac{E_c - F}{kT}}; \quad p_0 = N_v e^{-\frac{F - E_v}{kT}}; \tag{64.10}$$
$$n_0 p_0 = N_c N_v e^{-\frac{E_c - E_v}{kT}} = n_i^2.$$

If $r_l \neq g_l$, in adiabatic conditions the carrier concentration will vary. If $g_l > r_l$, $\frac{\partial n}{\partial t} > 0$, and the carrier concentration will increase with time; if $g_l < r_l$, $\frac{\partial n}{\partial t} < 0$, and it will decrease. Since the semiconductor is in adiabatic conditions and does not carry current, the condition $\frac{\partial n}{\partial t} \neq 0$ means that one of the processes is predominant. On the other hand, the absence of energy exchange and of current are the conditions necessary for the equilibrium state, and for this reason *the equation*

$$\frac{\partial n}{\partial t} = g_l - r_l \neq 0 \tag{64.11}$$

should describe a relaxation process, i.e. the process of establishing the state of equilibrium *disturbed by a current or by another external influence.* If the external influence resulted in an increase in carrier concentration ($n - n_0 = \delta n > 0$), recombination should dominate the relaxation process: $r_l > g_l$. If $n - n_0 = \delta n < 0$, thermal generation should be the dominant process: $g_l > r_l$. Denote the difference between g_l and r_l by R:

$$R = r_l - g_l. \tag{64.12}$$

The concentration variation rate for $r_l = r_0$ and $g_l = g_0$ is zero: $R = 0$, the solid is in the stationary equilibrium state. Speaking of recombination people usually have in mind the case $R \neq 0$. In other words, recombination is usually supposed to mean *excess carrier concentration variation*. To describe the relaxation process the dependence $R = R(\mathbf{r}, t)$ should be available. The continuity equation for this case,

$$\frac{\partial n}{\partial t} = -R(\mathbf{r}, t), \tag{64.13}$$

enables us to find the form of $n(\mathbf{r}, t)$.

Consider the solution of the equation (64.13) in certain assumptions concerning the value of $R(\mathbf{r}, t)$. Suppose there is *a constant quantity* $\frac{1}{\tau_f}$ *equal to the probability for a free carrier to recombine in unit time in unit volume*. In this case $\left(\frac{n-n_0}{\tau_f} = R\right)$ particles will recombine per unit time in unit volume:

$$\frac{\partial n}{\partial t} = -R = -\frac{n-n_0}{\tau_f} = -\frac{\delta n}{\tau_f}. \tag{64.14}$$

The equation (64.14) is easily integrated, and we obtain

$$\delta n(t) = n(t) - n_0 = [n(0) - n_0] e^{-\frac{t}{\tau_f}} = \delta n(0) e^{-\frac{t}{\tau_f}}. \tag{64.15}$$

Thus, *after the external influence has been removed, the non-equilibrium state relaxes with some characteristic parameter τ_f termed non-equilibrium carrier relaxation time or simply lifetime*. Numerically, τ_f is equal to the time during which the excess carrier concentration decreases e times. It may easily be seen that τ_f is *the mean lifetime of excess carrier concentration*.

Indeed, $\left(-dn = \delta n(t) \frac{dt}{\tau_f}\right)$ particles "aged" t will recombine during the interval $t, t+dt$. Should we add up all the ages $(-dn)\, t$ of the particles that have recombined and divide the sum by the initial number of charge carriers, we would obtain their average lifetime $\langle t \rangle$:

$$\langle t \rangle = \frac{1}{\delta n(0)} \int_0^\infty t(-dn) = -\int_0^\infty t e^{-\frac{t}{\tau_f}} \frac{dt}{\tau_f} = \tau_f. \tag{64.16}$$

Thus, the quantity τ_f, which determines the relaxation process and is equal to inverse recombination probability of one carrier per unit time per unit volume, is equal to the mean non-equilibrium carrier lifetime, or to the mean lifetime of non-equilibrium carrier concentration.

Since the equilibrium conditions may be disturbed both for electrons and for holes, the equations (64.16) and (64.15) are simultaneously valid for both types of carriers, although their lifetimes τ_f^n and τ_f^p need not coincide. *Recombination resulting in excess carrier concentration variation rates being proportional to their concentration is termed linear.* In case of linear recombination the transition of the carriers from the free state to the bound state takes place *independently of the availability of excess carriers of the opposite sign*. This means that *there is no direct electron-hole recombination.*

In case of direct recombination of excess electrons and holes one recombination act results in the annihilation of an electron and a hole. As long as we are dealing with direct electron-hole recombination, the concentration variation rates for the electrons and holes are equal. They should, moreover, be proportional to the product of their concentrations, or

$$R=\gamma\,(np-n_0p_0) \qquad (64.17)$$

and

$$\frac{\partial n}{\partial t}=-\gamma\,(np-n_0p_0)=\frac{\partial p}{\partial t}. \qquad (64.18)$$

Substituting $n_0+\delta n$ and $p_0+\delta p$ for n and p we obtain

$$\begin{aligned}\frac{\partial n}{\partial t}=\frac{\partial p}{\partial t}&=-\gamma\,[n_0p_0+n_0\delta p+p_0\delta n+\delta n\delta p-n_0p_0]=\\&=-\gamma\,[n_0\delta p+p_0\delta n+\delta n\delta p].\end{aligned} \qquad (64.19)$$

Let the deviation of the concentrations from their equilibrium values be small:

$$\delta n\delta p\ll n_0\delta p+p_0\delta n. \qquad (64.20)$$

Suppose the semiconductor is of n-type, so that $p_0\ll n_0$ and

$$\frac{\partial n}{\partial t}=\frac{\partial p}{\partial t}=-\gamma n_0\delta p. \qquad (64.21)$$

The solution of the equation (64.21) is

$$\delta p\,(t)=[p\,(0)-p_0]\,e^{-\gamma n_0 t}=\delta p\,(0)\,e^{-\frac{t}{\tau_f}}, \qquad (64.22)$$

where

$$\tau_f=\frac{1}{\gamma n_0}. \qquad (64.23)$$

It follows from (64.22) that the excess concentration of minority carriers decreases (or increases when $\delta p(0) < 0$) exponentially in the same way as in the case of linear recombination. This is quite natural since the equation (64.21) coincides with the equation (64.14) for $\gamma = (n_0\tau_f)^{-1}$.

Taking into account that

$$\frac{\partial n}{\partial t} = -\gamma n_0 \delta p(0) e^{-\frac{t}{\tau_f}} \qquad (64.24)$$

and integrating (64.24), we obtain

$$\delta n(t) = \gamma n_0 \delta p(0) \tau_f e^{-\frac{t}{\tau_f}} + C. \qquad (64.25)$$

The integration constant C may be determined from the initial conditions

$$\delta n(0) = \gamma n_0 \delta p(0)\, \tau_f + C = \delta p(0), \qquad (64.26)$$

or

$$C = \delta n(0) - \gamma n_0 \delta p(0)\, \tau_f = \delta n(0) - \delta p(0). \qquad (64.27)$$

Substituting (64.27) into (64.25) we obtain

$$\delta n(t) = \delta p(0) e^{-\frac{t}{\tau_f}} + \delta n(0) - \delta p(0). \qquad (64.28)$$

However, since $\delta n \to 0$ for $t \to \infty$, it follows that $\delta n(0) = \delta p(0)$ and $\delta n(t) = \delta p$, i.e. *in case of direct recombination the deviations of electron and hole concentrations from their respective equilibrium values are equal and vary exponentially with time, the exponent parameter being the lifetime* $\tau_f = (\gamma n_0)^{-1}$.

For high excess carrier concentrations $\delta n \gg (p_0, n_0)$

$$\frac{\partial n}{\partial t} = -\gamma \delta n \delta p. \qquad (64.29)$$

Presuming that $\delta n = \delta p$ we obtain

$$\frac{\partial \delta n}{\partial t} = \frac{\partial n}{\partial t} = -\gamma (\delta n)^2. \qquad (64.30)$$

Separating the variables we may write

$$\frac{\partial \delta n}{(\delta n)^2} = -\gamma\, dt, \qquad (64.31)$$

whence

$$\frac{1}{\delta n(t)} = \gamma t + C; \quad C = \frac{1}{\delta n(0)}, \qquad (64.32)$$

or

$$\delta n(t) = \frac{\delta n(0)}{1 + \gamma \delta n(0)\, t}. \qquad (64.33)$$

Hence, *in case of quadratic recombination, the excess carrier concentration decreases hyperbolically*. Should we calculate $\langle t\rangle$ we would obtain $\langle t\rangle=\infty$. This does not, however, mean that in case of quadratic recombination the non-equilibrium state persists for an infinitely long time. In some time, after which the condition $\delta n \gg (n_0, p_0)$ ceases to be valid, the hyperbolic law of δn variation changes into an exponential law.

If we agree that the lifetime concept may be introduced universally, in case of quadratic recombination the lifetime will depend on δp:

$$\frac{\partial \delta p}{\partial t}=-\gamma(\delta p)^2=-(\gamma\delta p)=-\frac{\delta p}{\tau_f(t)}, \tag{64.34}$$

where

$$\tau_f(t)=\frac{1}{\gamma\delta p(t)}. \tag{64.35}$$

The quantity $\tau_f(t)$ may be termed *instantaneous lifetime*. It is a function of excess carrier concentration.

Generally, it may be assumed that the variation of particle concentration is determined by the equation

$$\frac{\partial n}{\partial t}=-\frac{n-n_0}{\tau_f}, \tag{64.36}$$

where the relaxation lifetime is

$$\tau_f=\tau_f(t)=-\frac{n-n_0}{\frac{\partial n}{\partial t}}=-\frac{\delta n}{\frac{\partial \delta n}{\partial t}}. \tag{64.37}$$

In case of linear recombination this yields $\tau_f(t)=\tau_f$.

If the dependence $\delta n(t)$ is known, the instantaneous lifetime may be determined as the ratio of the $\delta n(t)$ curve ordinate to its derivative at the corresponding point. For linear recombination τ_f is equal to the time derivative of the curve drawn in the $\{\ln\delta n(t), t\}$ co-ordinates.

Consider the continuity equation for a uniform semiconductor in case of carrier generation induced by external excitation $G=g_E-r_E$ in the absence of electric current:

$$\frac{\partial n}{\partial t}=G-R. \tag{64.38}$$

We assume G to be independent of time. For a stationary state

$$G=R. \tag{64.39}$$

In case of linear recombination $R=\frac{n-n_0}{\tau_f}$ and

$$G=\frac{n-n_0}{\tau_f} \tag{64.40}$$

or

$$\delta n_{st} = G\tau_f \tag{64.41}$$

The quantity

$$\tau_f = \frac{\delta n_{st}}{G} = \tau_f^{st} \tag{64.42}$$

bears the name of stationary non-equilibrium carrier lifetime; it is equal to the ratio of stationary excess concentration value δn_{st} *to the generation rate* G.

For quadratic recombination

$$G = \gamma(\delta n_{st})^2; \quad \delta n_{st} = \sqrt{\frac{G}{\gamma}} \tag{64.43}$$

and

$$\tau^{st} = \frac{\delta n_{st}}{G} = \sqrt{\frac{1}{\gamma G}}. \tag{64.44}$$

For a non-stationary process

$$\frac{\partial n}{\partial t} = G - R \tag{64.45}$$

and

$$R = G - \frac{\partial n}{\partial t}. \tag{64.46}$$

Writing R in the form $R = \frac{n-n_0}{\tau_f}$ we may generally find the *instantaneous lifetime* from the relation

$$\tau_f = \tau_f(t) = \frac{n-n_0}{G - \frac{\partial n}{\partial t}}. \tag{64.47}$$

For $\frac{\partial n}{\partial t} = 0$, (64.47) defines the stationary lifetime; for $G = 0$, the relaxation lifetime.

If τ_f is constant, the equation (64.45) has a solution. Write

$$\frac{d(n-n_0)}{dt} = G - \frac{n-n_0}{\tau_f} = \frac{G\tau_f - (n-n_0)}{\tau_f}. \tag{64.48}$$

or

$$\frac{d(n-n_0)}{(n-n_0) - G\tau_f} = -\frac{dt}{\tau_f}.$$

Integrating (64.48) with the initial condition $\delta n(0) = 0$ we obtain

$$(n-n_0) - G\tau_f = Ce^{-\frac{t}{\tau_f}}; \quad C = -G\tau_f;$$

$$\delta n(t) = G\tau_f\left(1 - e^{-\frac{t}{\tau_f}}\right). \tag{64.49}$$

Should the excitation be switched off at $t \gg \tau_f$, it would follow

$$\delta n(t) = \delta n(0) e^{-\frac{t}{\tau_f}} = G\tau_f e^{-\frac{t}{\tau_f}}. \qquad (64.50)$$

Figure 95 shows the dependence $\delta n(t)$ for the case of a semiconductor excited by periodic light pulses G, the duration of the light and dark periods being $T \gg \tau_f$.

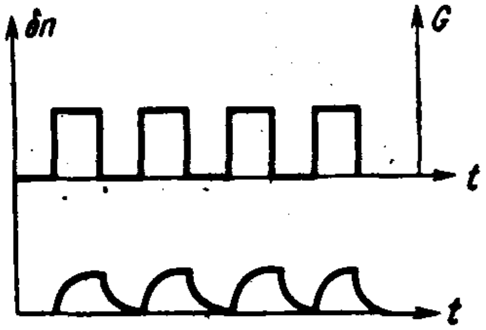


Fig. 95. The variation with time of excess carrier concentration δn created by illumination of the semiconductor with rectangular light pulses

In case of quadratic recombination the continuity equation assumes the form

$$\frac{\partial n}{\partial t} = \frac{\partial \delta n}{\partial t} = G - \gamma (\delta n)^2. \qquad (64.51)$$

Separate the variables:

$$\frac{d(n-n_0)}{G-\gamma(n-n_0)^2} = dt, \qquad (64.52)$$

whence

$$t = \int \frac{d(n-n_0)}{G-\gamma(n-n_0)^2} + C. \qquad (64.53)$$

The integral is computed as follows: find the roots of the denominator

$$G-\gamma(\delta n)^2 = 0, \quad \delta n_{1,2} = \pm \sqrt{\frac{G}{\gamma}} = \pm \delta n_1. \qquad (64.54)$$

Factorize the integrand function:

$$\frac{1}{G-\gamma(\delta n)^2} = \frac{1}{(\delta n_1 - \delta n)(\delta n_1 + \delta n)\gamma} = \frac{1}{\gamma}\left(\frac{A}{\delta n_1 - \delta n} + \frac{B}{\delta n_1 + \delta n}\right) = $$
$$= \frac{1}{\gamma}\frac{\delta n_1(A+B) + \delta n(A-B)}{(\delta n_1 - \delta n)(\delta n_1 + \delta n)}, \qquad (64.55)$$

whence

$$A = B, \quad A = \frac{1}{2\delta n_1} = \frac{1}{2}\sqrt{\frac{\gamma}{G}}. \qquad (64.56)$$

Taking into account (64.56) and (64.55) we obtain for (64.53)

$$t=\frac{1}{2\sqrt{G\gamma}}\ln\frac{\delta n_1+\delta n}{\delta n_1-\delta n}+C. \qquad (64.57)$$

The condition $\delta n(0)=0$ yields $C=0$ or

$$\frac{\delta n_1+\delta n}{\delta n_1-\delta n}=e^{2\sqrt{G\gamma}\,t}, \qquad (64.58)$$

whence

$$\delta n(t)=\delta n_1\frac{e^{2\sqrt{G\gamma}t}-1}{e^{2\sqrt{G\gamma}t}+1}=\sqrt{\frac{G}{\gamma}}\tanh\sqrt{G\gamma}\,t. \qquad (64.59)$$

Thus, *in case of quadratic recombination the increase in excess carrier concentration is determined by a hyperbolic tangent curve.* For small times $t\ll\frac{1}{\sqrt{G\gamma}}$ we have

$$\delta n(t)\cong\sqrt{\frac{G}{\gamma}}\sqrt{G\gamma}t=Gt. \qquad (64.60)$$

For large times $t\gg\frac{1}{\sqrt{G\gamma}}$

$$\delta n(t)\cong\sqrt{\frac{G}{\gamma}}=\delta n_{st}. \qquad (64.61)$$

If a stationary excitation is switched off in a definite moment of time, the excess carrier concentration will decrease hyperbolically:

$$\delta n(t)=\frac{\delta n(0)}{1+\sqrt{G\gamma}t}=\frac{\sqrt{\frac{G}{\gamma}}}{1+\sqrt{G\gamma}\,t}. \qquad (64.62)$$

The smaller is γ, the more rapid is the rise in excess carrier concentration after the excitation had been switched on and the more slowly it attenuates after the excitation has been switched off.

Summary of Sec. 64

1. A state other than the state of thermodynamic equilibrium is termed non-equilibrium state, and charge carriers in this state are termed non-equilibrium. The non-equilibrium state may be brought about by the re-distribution of carriers inside the Brillouin zone with the total carrier concentration in the zone remaining constant. Such a state exists, for instance, when electric current flows in a uniform semiconductor. A second type of non-equilibrium states sets in when changes in carrier concentration are caused by internal or external factors. The particle concentration (of ele-

ctrons or holes) is in this case termed non-equilibrium, and the difference

$$n(\mathbf{r},\ t)-n_0(\mathbf{r})=\delta n(\mathbf{r},\ t), \tag{64.1s}$$

excess carrier concentration. In semiconductor and semiconductor-device physics the term non-equilibrium state is usually applied to the case $\delta n(\mathbf{r},\ t)\neq 0$. The return to the equilibrium state in the first instance proceeds with a characteristic time-constant τ termed relaxation time which had been used in the description of kinetic effects. In the second instance this characteristic time-constant is termed lifetime and is denoted by τ_f (frequently the notation used for lifetime is τ; as a rule, this does not lead to misunderstanding, and below this will be the notation we shall use for the lifetime). The meaning of the lifetime concept depends on specific conditions in which the semiconductor finds itself.

2. Continuity equation in the form

$$\frac{\partial n}{\partial t}=-\operatorname{div}\frac{\mathbf{j}}{e}+g_I+g_E-r_I-r_E, \tag{64.2s}$$

where $n(\mathbf{r},\ t)$ is the particle concentration; $\frac{\mathbf{j}}{e}$—their flow density; g_I, g_E—generation rates; r_I, r_E—recombination rates determined by internal $(g_I,\ r_I)$ and external $(g_E,\ r_E)$ causes, may be used to describe the non-equilibrium state. Two quantities, R and G,

$$\mathrm{R}=r_I-g_I;\quad \mathrm{G}=g_E-r_E, \tag{64.3s}$$

may be used instead of the four quantities r_I, g_I, r_E, g_E. R is termed recombination rate, and G—generation rate. The continuity equation may be written in the form

$$\frac{\partial n}{\partial t}=-\operatorname{div}\frac{\mathbf{j}}{e}+\mathrm{G}-\mathrm{R}. \tag{64.4s}$$

3. Instantaneous lifetime τ_f is the term applied to the ratio of excess concentration to recombination rate in the absence of current:

$$\frac{\delta n}{\tau_f}=\mathrm{R}=\mathrm{G}-\frac{\partial n}{\partial t}, \tag{64.5s}$$

or

$$\tau_f=\frac{n-n_0}{\mathrm{R}}=\frac{n-n_0}{\mathrm{G}-\frac{\partial n}{\partial t}}. \tag{64.6s}$$

4. In a stationary state $\frac{\partial n}{\partial t}=0$; $n-n_0=\delta n_{st}$, the stationary lifetime being determined by (64 7s):

$$\tau_f^{st}=\frac{\delta n_{st}}{\mathrm{G}};\quad \delta n_{st}=\mathrm{G}\tau_f^{st}. \tag{64.7s}$$

5. Recombination is termed linear if the recombination rate R is proportional to the excess concentration of carriers of the same sign: $R = \frac{\delta n}{\tau_f}$; for $\tau_f = \text{const}$ this yields

$$\frac{\partial n}{\partial t} = -\frac{\delta n}{\tau_f} + G. \tag{64.8s}$$

For $G = 0$ (no generation)

$$\delta n(t) = \delta n(0)\, e^{-\frac{t}{\tau_f}}. \tag{64.9s}$$

For $G \neq 0$ and $\delta n(0) = 0$ the excess carrier concentration tends to a saturation value

$$\delta n(t) = G\tau_f\left(1 - e^{-\frac{t}{\tau_f}}\right). \tag{64.10s}$$

6. In case of quadratic recombination $R = \gamma(np - n_0 p_0)$, and excess concentration decreases hyperbolically with time. The increase of excess concentration with time follows a hyperbolic tangent curve (64.62) and (64.59).

65. RECOMBINATION MECHANISM. LINEAR RECOMBINATION

The non-equilibrium charge carrier state may be described with the aid of some distribution function $f(\mathbf{r}, \mathbf{k}, t)$ which was obtained in Chapter IV for the case when the deviation from the equilibrium state was caused by the fields **E** and **B**, by chemical potential gradients F (gradients of Fermi energy) and by temperature gradients. The non-equilibrium part $f^{(1)}$ of the distribution function describes the re-distribution of particles among the states of the Brillouin zone with the number of particles remaining constant. The nature of the changes in the distribution function in the case of non-equilibrium states of the second type, when particles concentration varies, should be different. We presume that we may use the Fermi-Dirac distribution function with modified parameters to describe the non-equilibrium states.

Systems in which the probability of finding a particle in higher states is greater than that in lower states are termed inversely populated systems. To describe them their temperature should be assumed *negative*. Negative-temperature systems are used for quantum generators and amplifiers.

The function used to describe non-equilibrium systems with normal distribution of particles among states is the Fermi-Dirac function (or Boltzmann function) in which a definite quantity F^*, termed quasi-Fermi level, is substituted for the Fermi energy F. The quasi-Fermi level F^* may be introduced as a quantity which describes normal distribution of particles among states in systems with non-equilibrium carrier concentrations. The value of F^* may be determined from the normalizing condition in the same way it was done for the Fermi level:

$$n = \frac{2N_c}{\sqrt{\pi}} \Phi_{1/2}(\xi^*), \quad \xi^* = \frac{F^* - E_c}{kT}; \tag{65.1}$$

$$p = \frac{2N_v}{\sqrt{\pi}} \Phi_{1/2}(\eta^*), \quad \eta^* = \frac{E_v - F^*}{kT}. \tag{65.2}$$

We may write for non-degenerate semiconductors

$$n = N_c e^{-\frac{E_c - F_n^*}{kT}}; \quad p = N_v e^{-\frac{F_p^* - E_v}{kT}} \tag{65.3}$$

and

$$np = N_c N_v e^{-\frac{\Delta E_g}{kT}} e^{\frac{F_n^* - F_p^*}{kT}} = n_i^2 e^{\frac{F_n^* - F_p^*}{kT}} \tag{65.4}$$

As may be seen from (65.4), in non-equilibrium systems the quasi-Fermi levels of electrons and holes do not coincide. *The greater is the difference between the quasi-Fermi levels of electrons and holes, the greater is the difference between the products of their non-equilibrium and equilibrium concentrations.*

Before we employ the quasi-Fermi level concept to describe recombination processes in non-equilibrium systems consider some models of recombination mechanisms.

The recombination mechanisms are classified into three types *according to the method of energy exchange.*

1. The term radiative, or photon, recombination applies to recombination in the course of which the energy of recombining particles is radiated in the form of a photon.

2. If the energy of the particle is transmitted to the lattice (to the phonons) the recombination is termed non-radiative, or phonon.

3. In one of the non-radiative recombination mechanisms, the impact ionization (Auger processes), the energy of recombining particles is transmitted to a third particle which because of this becomes "hot". The "hot" particle in multiple collisions gives up its energy to the phonons.

In addition to these three main mechanisms the energy of recombining particles may be transmitted to the electron gas (plasma recombination). If the electron and hole form an exciton as an intermediate stage, such a recombination is termed exciton recombination.

The photon, phonon, and Auger recombination may take a different course depending on *the mechanism of electron transition* from the conduction to the valence band.

If the electron and the hole recombine as a result of a direct collision the recombination is termed *direct*, or *interband*. Direct recombination plays the leading role in solids of narrow forbidden band width of the order of 0.2-0.3 eV.

For forbidden band widths in excess of 0.5 eV the recombination proceeds via *localized states* inside the forbidden band. These states are usually termed *recombination traps*. Suppose the semiconductor contains a concentration N_t of defects whose energy levels E_t lie inside the forbidden band and are not occupied by electrons (are occupied by holes). In this case several processes illustrated in Fig. 96 are possible:

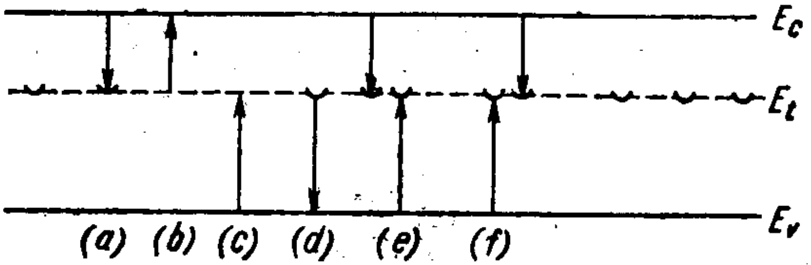


Fig. 96. A schematic diagram of electron and hole transitions due to the interaction of traps and recombination centres with the energy bands

(a) a free electron is trapped by a neutral defect;

(b) an electron from a negatively charged defect goes over to the conduction band. In this way a free electron, after spending some time trapped on the defect level, becomes free again. A defect with the energy level E_t that traps electrons and subsequently sets them free is termed *electron trap*;

(c) a free hole is trapped by a neutral defect (an electron from the defect goes over to the valence band);

(d) an electron from the valence band is trapped by a positively charged defect. Such a defect acts as a *hole trap*;

(e) by trapping an electron from the conduction band a neutral defect becomes negatively charged and traps a free hole, i.e. transfers the trapped electron to the valence band. The result is a recombination process of the electron-hole pair;

(f) by trapping a free hole a neutral defect becomes positively charged and traps a free electron, again becoming neutral. The result is a recombination process of the electron-hole pair.

Charge carrier trapping does not affect stationary lifetime but affects instantaneous lifetime values. The liberation of the charged carrier may be caused by thermal transfer. Sometimes it is the result of "illumination".

Carrier trapping by a neutral defect enhances the probability of the subsequent trapping of an oppositely charged carrier owing to Coulomb attraction forces between the charged recombination trap and the charge carrier of opposite sign.

Thus, *defects with the energy level E_t can act as traps for electrons and holes.* Donor and acceptor atoms, too, can act as traps. Characteristic of the traps is that they interact mainly with one band—with the conduction or the valence band. To provide a recombination centre for an electron-hole pair the defect should effectively interact with both bands.

Let us now discuss the recombination process via recombination traps whose energy level is E_t and the concentration N_t. Since the semiconductor is in a non-equilibrium state, electron and hole distribution over states in energy bands is described by the quasi-Fermi levels F_n^*, F_p^*, F_t^*. Denote the electron and hole distribution functions for the defect level E_t by f_t and $f_{tp}=1-f_t$, respectively.

Assume the existence of *the probabilities per unit time for a neutral defect to trap a free electron, and for a negatively charged defect to trap a free hole* and denote them by $c_n(E)$ and $c_p(E)$.

The energy interval dE contains $2N(E)\,dE$ states accommodating $dn_c=2N(E)\,f\,dE$ electrons. The number of electrons trapped per unit time by the recombination traps will be

$$dr_n=c_n(E)\,2N(E)\,f\,dEN_tf_{tp}. \qquad (65.5)$$

During the same time

$$dg_n=l_n(E)\,2N(E)\,f_p\,dEN_tf_t \qquad (65.6)$$

electrons will go over to the conduction band. The quantity $l_n(E)$ denotes the probability per unit time for an electron to go over from the trap level to a free level E in the conduction band. A definite relationship should exist between $c_n(E)$ and $l_n(E)$ which may be found from the thermodynamic equilibrium condition. In this condition

$$dr_n=dg_n. \qquad (65.7)$$

Substituting (65.6) and (65.5) into (65.7) and cancelling out similar terms we obtain

$$c_n(E)\,ff_{tp}=l_n(E)\,f_pf_t, \qquad (65.8)$$

whence

$$\frac{l_n(E)}{c_n(E)}=\frac{f}{f_p}\frac{f_{tp}}{f_t}. \qquad (65.9)$$

However, since

$$f_{tp}=1-f_t=1-\frac{1}{e^{\frac{E_t-F_t^*}{kT}}+1}=f_t e^{\frac{E_t-F_t^*}{kT}}, \qquad (65.10)$$

and

$$f_p=fe^{\frac{E-F_n^*}{kT}}, \qquad (65.11)$$

it follows

$$\frac{l_n(E)}{c_n(E)}=e^{-\frac{E-F_n^*}{kT}}e^{\frac{E_t-F_t^*}{kT}}=e^{-\frac{E-E_t}{kT}}, \qquad (65.12)$$

since in conditions of equilibrium

$$F_t^*=F_n^*=F_p^*=F. \qquad (65.13)$$

In the absence of equilibrium $dg_n \neq dr_n$, therefore

$$\begin{aligned} dR_n &= dr_n-dg_n=c_n(E)\,2N(E)\,f\,dEN_tf_{tp}-l_n(E)\,2N(E)\,f_p dE\,N_tf_t= \\ &= c_n(E)\,2N(E)\,dE\,fN_tf_{tp}\left[1-\frac{l_n(E)}{c_n(E)}\frac{f_p}{f}\frac{f_t}{f_{tp}}\right]= \\ &= c_n(E)\,2N(E)\,dE\,fN_tf_{tp}\left[1-e^{\frac{F_t^*-F_n^*}{kT}}\right]. \end{aligned} \qquad (65.14)$$

Integrating dR_n over energy we obtain

$$R_n=N_tf_{tp}\left[1-e^{\frac{F_t^*-F_n^*}{kT}}\right]2\int_{E_c}^{\infty}N(E)fc_n(E)dE=\left[1-e^{\frac{F_t^*-F_n^*}{kT}}\right]N_tf_{tp}n\langle c_n\rangle. \qquad (65.15)$$

Transform the expression for R_n as follows. Denote

$$N_t\langle c_n\rangle=C_n=\frac{1}{\tau_{0n}}, \qquad (65.16)$$

and re-write the second addend in (65.15) (without C_n):

$$ne^{\frac{F_t^*-F_n^*}{kT}}f_{tp}=N_ce^{\frac{-E_c+F_n^*+F_t^*-F_n^*+E_t-F_t^*}{kT}}f_t=N_ce^{-\frac{E_c-E_t}{kT}}f_t=n_1f_t, \qquad (65.17)$$

where n_1 is numerically equal to the electron concentration when the Fermi level F_n^* coincides with the trap level E_t. Taking into account (65.17) we write the electron recombination rate R_n in the form

$$R_n = C_n n f_{tp} - C_n n_1 f_t. \tag{65.18}$$

Find the rate of hole concentration variation. We will assume the existence of *the probabilities per unit time for a hole to be trapped* $c_p(E)$, *and to be emitted by a trap*, $l_p(E)$. Repeating the above reasoning for the holes we obtain

$$R_p = C_p p f_t - C_p p_1 f_{tp};\quad \frac{1}{C_p} = \tau_{p0}, \tag{65.19}$$

where

$$p_1 = N_v e^{-\frac{E_t - E_v}{kT}} \tag{65.20}$$

The combined trapping rates of electrons and holes by the recombination centres are equal:

$$R_p = R_n. \tag{65.21}$$

This will enable us to find the function f_t, i.e. *the position of the level* F_t^*:

$$C_p p f_t - C_p p_1 f_{tp} = C_n n f_{tp} - C_n n_1 f_t, \tag{65.22}$$

or

$$(C_p p + C_n n_1) f_t = (C_n n + C_p p_1)(1 - f_t), \tag{65.23}$$

whence

$$f_t = \frac{C_n n + C_p p_1}{C_n(n + n_1) + C_p(p + p_1)} \tag{65.24}$$

and

$$f_{tp} = 1 - f_t = \frac{C_n n_1 + C_p p}{C_n(n + n_1) + C_p(p + p_1)}. \tag{65.25}$$

Substituting f and f_{tp} into (65.18) we obtain

$$R_n = \frac{C_n n(C_n n_1 + C_p p) - C_n n_1(C_n n + C_p p_1)}{C_n(n + n_1) + C_p(p + p_1)} = \frac{C_n C_p(np - n_1 p_1)}{C_n(n + n_1) + C_p(p + p_1)}. \tag{65.26}$$

But according to (65.4)

$$np = n_i^2 e^{\frac{F_n^* - F_p^*}{kT}}, \tag{65.27}$$

and it follows from (65.17) and (65.20) that

$$n_1 p_1 = N_c e^{-\frac{E_c - E_t}{kT}} N_v e^{-\frac{E_t - E_v}{kT}} = n_i^2, \tag{65.28}$$

therefore

$$R_n=\frac{C_nC_p\left[np-n_i^2\right]}{C_n(n+n_1)+C_p(p+p_1)}=\frac{C_nC_pn_i^2\left[e^{\frac{F_n^*-F_p^*}{kT}}-1\right]}{C_n(n+n_1)+C_p(p+p_1)}. \qquad (65.29)$$

Define the electron τ_n and hole τ_p lifetime by the relation (64.6s):

$$\tau_n=\tau_f^n=\frac{n-n_0}{R_n}=\frac{\delta n}{R_n}=\frac{\delta n\,[C_n(n+n_1)+C_p(p+p_1)]}{C_nC_p\left[np-n_i^2\right]}, \qquad (65.30)$$

$$\tau_p=\tau_f^p=\frac{\delta p}{R_p}. \qquad (65.31)$$

Since $R_n=R_p$ it follows that $\delta n=\delta p$ and $\tau_n=\tau_p=\tau_f$, where τ_f is the electron-hole pair lifetime.

Setting $n=n_0+\delta n$, $p=p_0+\delta p$ we obtain

$$np-n_0p_0=n_0p_0+\delta np_0+\delta pn_0+\delta n\delta p-n_0p_0=\delta n\,(n_0+p_0+\delta n) \qquad (65.32)$$

and

$$\tau_n=\tau_{p0}\frac{n_0+n_1+\delta n}{n_0+p_0+\delta n}+\tau_{n0}\frac{p_0+p_1+\delta p}{n_0+p_0+\delta p}. \qquad (65.33)$$

Analyse the expression (65.33) for the electron-hole pair lifetime. The non-equilibrium charge carrier lifetime is determined by two terms related to the trapping times of the free carriers by the defect, τ_{p0} and τ_{n0}, respectively. *In addition to the trapping probability which depends on the nature of the traps and on their concentration N_t, the lifetime depends also on the excess carrier concentrations δn and δp.* The position of the level E_t is determined by n_1 and p_1. The donor and acceptor concentrations enter the expression for the lifetime via the equilibrium charge carrier concentrations n_0 and p_0. Hence, the following factors affect non-equilibrium carrier lifetime:

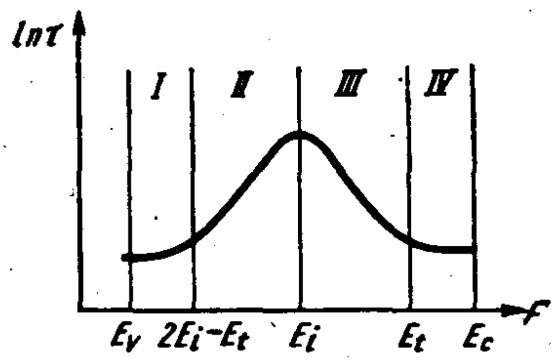


Fig. 97. The dependence of ln τ on the position of the Fermi level inside the forbidden band

(1) recombination trap type and concentration (via the times τ_{p0} and τ_{n0});

(2) doping impurity concentration (via the concentrations n_0 and p_0);

(3) the position of the defect level E_t (via n_1 and p_1);

(4) excess non-equilibrium carrier concentrations (δn, δp);

(5) temperature (determines the values of n_0, p_0, n_1, p_1 and, possibly, of τ_{n0}, τ_{p0} as well).

Figure 97 shows the position of the defect energy level E_t in the forbidden band. To avoid ambiguity, E_t may be assumed to lie between the middle of the forbidden band E_i and the bottom of the conduction band E_c; in this case $n_1 > n_i > p_1$. Divide the forbidden band symbolically into four regions.

Consider first the case of *small deviations of carrier concentration from its equilibrium value*:

$$\delta n = \delta p \ll (p_0,\ n_0). \tag{65.34}$$

The expression for the lifetime of a pair is of the form

$$\tau = \tau_{p0}\frac{n_0+n_1}{n_0+p_0} + \tau_{n0}\frac{p_0+p_1}{n_0+p_0}. \tag{65.35}$$

Consider the dependence of τ on the equilibrium concentrations n_0, p_0 determined by the Fermi level position F.

1. **The Fermi level lies in region I:**

$$E_v < F < 2E_i - E_t = E_v + E_c - E_t; \tag{65.36}$$

the semiconductor is of the hole type: $p_0 \gg n_0$; besides, $p_0 \gg p_1$; $p_0 \gg n_1$.

The equation (65.35) may be simplified:

$$\tau = \tau_{p0}\frac{n_0+n_1}{n_0+p_0} + \tau_{n0}\frac{p_0+p_1}{n_0+p_0} \cong \tau_{n0}. \tag{65.37}$$

The lifetime of a pair of non-equilibrium carriers is determined solely by the trapping time τ_{n0} of the electrons, the minority carriers in p-type simiconductor. Immediately as minority carrier is trapped by the recombination centre a majority carrier is trapped, since the concentration of the latter is great.

2. **The Fermi level lies in region IV:**

$$E_t < F < E_c; \quad n_0 \gg n_1 \gg p_1 \gg p_0. \tag{65.38}$$

Simplifying the expression (65.35) we obtain, as in case 1,

$$\tau \cong \tau_{p0}. \tag{65.39}$$

Thus, non-equilibrium charge carrier lifetime is determined solely by the nature of the recombination centres and does not depend on the minority and majority carrier concentrations.

3. **The Fermi level lies in region II;** $E_c - E_t < F < E_i$; the semiconductor is of lightly doped p-type; the nearer is F to E_i, the closer are the concentrations p_0 and n_0 to the intrinsic concentration. Taking into account the relations existing between the quantities contained in (65.33), $n_1 \gg p_0 \gg n_i \gg n_0 \gg p_1$, we obtain

$$\tau \cong \tau_{n0} + \tau_{p0}\frac{n_1}{p_0}. \tag{65.40}$$

For $F=E_i$ the semiconductor is intrinisic and $\tau \cong \tau_{p0}\frac{n_1}{n_i}$. For $F \cong E_c - E_t$, $p_0 \cong n_1$ and $\tau \cong \tau_{p0}+\tau_{n0}$. Hence, as F moves from the left border to the right border of region II, the lifetime τ increases from $\tau_{n0}+\tau_{p0}$ to $\tau_{p0}\frac{n_1}{n_i} \gg \tau_{n0}+\tau_{p0}$.

4. **The Fermi level lies in region III:** $E_i < F < E_t$; $n_1 \gg n_0 \gg n_i \gg p_0 \gg p_1$; therefore, it follows from (65.35) that

$$\tau \cong \tau_{n0} + \tau_{p0}\frac{n_1}{n_0}. \tag{65.41}$$

As F moves from E_i to E_t and enters region IV, τ changes from $\tau_{p0}\frac{n_1}{n_i}$ to $\tau \cong \tau_{p0}$.

Figure 97 shows in graphical form the dependence of $\ln\tau$ on $\ln\frac{n}{n_i}$. *The lifetime τ_i is at its maximum in an intrinsic semiconductor, being equal to the minority carrier trapping time in the heavily doped semiconductor.*

If E_t lies below E_i the nature of the dependence of the lifetime on n_0 and p_0 will remain unchanged, except that in the regions II and III the lifetime will be determined by the quantity $\tau_{n0}\frac{p_1}{n_0}$ or $\tau_{n0}\frac{p_1}{p_0}$. The nearer is E_t to E_i, the narrower is the range of concentrations with high lifetime values. In case of a semiconductor with a non-uniform impurity distribution the lifetimes will be less in the heavier doped parts (a uniform distribution of recombination centres is presumed), the greatest lifetimes being in the *p-n* junction region.

Consider now the case of a large deviation of charge carrier concentrations from their equilibrium values (the so-called *large injection level*): $n \gg n_0$; $p \gg p_0$.

Transform the expression for τ so that the low injection level quantity $\tau = \tau_0$ could be separated:

$$\tau = \tau_{p0}\frac{n_0+n_1+\delta n}{n_0+p_0+\delta n} + \tau_{n0}\frac{p_0+p_1+\delta p}{n_0+p_0+\delta p} =$$

$$= \frac{\tau_{p0}\frac{n_0+n_1}{n_0+p_0} + \tau_{n0}\frac{p_0+p_1}{n_0+p_0} + (\tau_{p0}+\tau_{n0})\frac{\delta n}{n_0+p_0}}{1+\frac{\delta n}{n_0+p_0}} = \tau_0\frac{1+a\delta n}{1+b\delta n}, \tag{65.42}$$

where

$$\tau_0 = \tau_{p0}\frac{n_0+n_1}{n_0+p_0} + \tau_{n0}\frac{p_0+p_1}{n_0+p_0}; \tag{65.43}$$

$$a = \frac{\tau_{p0}+\tau_{n0}}{\tau_{p0}(n_0+n_1)+\tau_{n0}(p_0+p_1)}; \quad b = \frac{1}{n_0+p_0}.$$

The quantities a and b are positive. Denote the lifetime for $\delta n \longrightarrow \infty$ by τ_∞. It follows from (65.42) that

$$\tau_\infty = \tau_0 \frac{a}{b}. \tag{65.44}$$

Consider the ratio

$$\frac{a}{b} = \frac{(\tau_{p0} + \tau_{n0})(n_0 + p_0)}{\tau_{p0}(n_0 + n_1) + \tau_{n0}(p_0 + p_1)}, \tag{65.45}$$

or

$$\frac{b}{a} = \frac{\tau_{p0}}{\tau_{p0} + \tau_{n0}} \frac{n_0 + n_1}{n_0 + p_0} + \frac{\tau_{n0}}{\tau_{p0} + \tau_{n0}} \frac{p_0 + p_1}{p_0 + n_0} = \frac{\tau_0}{\tau_{p0} + \tau_{n0}} \tag{65.46}$$

and

$$\tau_\infty = \tau_{p0} + \tau_{n0}. \tag{65.47}$$

Since τ_0 may be both less and greater than $(\tau_{p0} + \tau_{n0})$, $\frac{a}{b}$, too, may be both greater and less than unity. Therefore, depending on the position of the Fermi level relative to E_t, τ may either be greater than τ_0, or less. For $a > b$ the lifetime τ rises with the rise in the non-equilibrium charge carrier injection level, the opposite being true for $a < b$.

The Hall-Shockley-Read recombination theory discussed in this section applies to a model of the simplest kind and is valid for low trap concentrations N_t. Real semiconductors may contain recombination traps of different nature, therefore the lifetime variations may be much more complicated. Lifetimes in germanium and silicon are greatly affected by atomic impurities such as nickel, manganese, and gold; an important part is played by dislocations, vacancies and interstitials.

Lifetimes in germanium and silicon depend on their composition and may vary from miliseconds to fractions of a microsecond. In other semiconductors the lifetimes may be as great as several seconds and as low as 10^{-16} s and even lower.

In an intrinsic semiconductor the lifetime of a carrier pair is, in essence, the minority carrier trapping time; for instance in an electron-type semiconductor the lifetime τ_f is determined by the hole trapping time τ_{p0}. The trapping time, however, may be, in compliance with (65.16), expressed in terms of the recombination trap concentration N_t:

$$\tau_{p0} = \frac{1}{C_p} = \frac{1}{\langle c_p \rangle N_t}. \tag{65.48}$$

Should we, instead of the trapping probability by one trap of one carrier per unit volume per unit time, introduce the trapping probability of one charge carrier from a unit particle flow, we would be able to arrive at the definition of effective scattering cross section. By changing the type of traps and their concentration one may change the non-equilibrium carrier trapping time and, thereby, the lifetime of a carrier pair. Table 21 presents the properties of some impurities in germanium and silicon which serve as recombination centres.

Table 21

Impurity or defect	Germanium			Silicon		
	E_t, eV	$\langle c_p\rangle\cdot 10^9$, cm^3·s	$\langle c_n\rangle\cdot 10^9$, cm^3·s	E_t, eV	$\langle c_p\rangle\cdot 10^9$, cm^3·s	$\langle c_n\rangle\cdot 10^9$, cm^3·s
Copper	+0,31	—	0.25	+0.24	—	—
Silver	+0.13 —0.28	— 100	1.5 2.5	— —	— —	— —
Gold	+0.16 —0.20	1000 200	2.5 5	+0.35 —0.54	1.6 16	67 9.5
Gallium	+0.0108	—	2×10^{-3}	+0.065	—	10^{-3}
Iron	+0.35 —0.27	60 200	25 25	+0.40 —0.55	— 2	— —
Nickel	+0.23 —0.30	— 200	2.5 7.5	—	—	—
Vacancies	+0.36	7	—	—	—	—
Interstitials	+0.42	0.14	—	—	—	—

The position of the energy level is shown with a plus sign if E_t is measured from E_v, and with a minus sign if it is measured from E_c. The electron trapping probability of a gallium atom is shown for comparison.

Summary of Sec. 65

1. A defect capable of trapping free charge carriers and of subsequently releasing them is termed a trap.

2. A defect capable of trapping subsequently an electron and a hole, the result being the recombination of a pair of free carriers, is termed recombination trap or centre.

3. The theory of electron and hole recombination via recombination centres arrives at the following expression for the lifetime of a pair $\tau \cong \tau_f$:

$$\tau_f = \tau_{p0} \frac{n_0 + n_1 + \delta n}{n_0 + p_0 + \delta n} + \tau_{n0} \frac{p_0 + p_1 + \delta p}{n_0 + p_0 + \delta p}. \qquad (65.1s)$$

This expression describes the dependence of the lifetime of a pair on: (a) excess non-equilibrium carrier concentration $\delta n = \delta p$ (on the injection level); (b) on the equilibrium carrier concentrations n_0, p_0; (c) on the position of the recombination centre energy level E_t (via n_1 and p_1); (d) on the defect concentration N_t and on the trapping probabilities $\langle c_n \rangle$ and $\langle c_p \rangle$ which determine the trapping times τ_{n0} and τ_{p0}.

4. For a low injection level in the impurity conductivity range the lifetime of a pair is determined by the minority carrier trapping time. As the Fermi level moves towards the middle of the forbidden band, the lifetime of a pair increases.

5. The Fermi-Dirac function with modified parameters may be used to describe non-equilibrium states. If the upper levels are more densely populated than the lower levels, the temperature of the system of non-equilibrium particles must be assumed to be negative: $T < 0$. Such systems are termed systems with inversely populated levels, or negative-temperature systems; they are employed in quantum oscillators and amplifiers.

A system with a normal distribution of non-equilibrium particles over energy levels is described by the quasi-Fermi level F^* determined from the condition of normalizing the non-equilibrium distribution function to the non-equilibrium concentration n:

$$n = \frac{2N_c}{\sqrt{\pi}} \Phi_{1/2}(\xi^*). \qquad (65.2s)$$

For non-degenerate semiconductors the position of the quasi-Fermi level may be found for electrons from the equation

$$n = N_c e^{-\frac{E_c - F_n^*}{kT}}; \quad F_n^* = E_c + kT \ln \frac{n}{N_c}, \qquad (65.3s)$$

and for holes, from the equation

$$p = N_v e^{-\frac{F_p^* - E_v}{kT}}; \quad F_p^* = E_v - kT \ln \frac{p}{N_v}. \qquad (65.4s)$$

The separation of the quasi-Fermi levels for electrons and holes is determined by the injection level:

$$F_n^* - F_p^* = kT \ln \frac{np}{n_0 p_0} = kT \ln \frac{np}{n_i^2}. \qquad (65.5s)$$

For degenerate semiconductors the relation between the non-equilibrium carrier concentration and the quasi-Fermi level follows from the expression

$$\Phi_{1/2}(\xi^*) = \frac{2}{3} \xi^{*3/2}.$$

66. DIFFUSION AND DRIFT OF NON-EQUILIBRIUM CHARGE CARRIERS

Several new phenomena may be observed in semiconductors containing non-equilibrium charge carriers. These phenomena may be described on the basis of the generalized Boltzmann equation or with the aid of the continuity equation which follows from it:

$$\frac{\partial n}{\partial t} = -\operatorname{div} \frac{\mathbf{j}}{e} + \mathrm{G} - \mathrm{R}. \qquad (66.1)$$

Consider a stationary state in the absence of charge carrier generation caused by external sources of energy, i.e. set $\frac{\partial n}{\partial t} = 0$, $\mathrm{G} = 0$, and write the equation (66.1) in the form

$$\frac{1}{e} \operatorname{div} \mathbf{j} + \mathrm{R} = 0. \qquad (66.2)$$

It follows from the equation (66.2) that for $\mathrm{R} \neq 0$ the non-equilibrium charge carrier concentration is sustained by the current $\mathbf{j}$. In other words, *the decrease in the excess carrier concentration due to recombination is compensated by the divergence of the current.*

We may use the general expression (38.1s) for the current density $\mathbf{j}$ assuming the kinetic coefficients to be expressed in terms of non-equilibrium concentration n and averaged relaxation time of non-equilibrium charge carriers. Suppose that there is no magnetic field ($\mathbf{B} = 0$) and that the semiconductor is in isothermal conditions ($\nabla T = 0$). In this case we may write, following (38.1s),

$$\mathbf{j} = e^2 K_{11} \mathbf{E} - e K_{11} \nabla F^*. \qquad (66.3)$$

It is obvious that the term $-eK_{11}\nabla F^*$ should be retained since it describes the current due to the inhomogeneity of the semiconductor. In this case the cause of the inhomogeneity is a possible co-ordinate dependence of non-equilibrium charge carrier concentration. Re-write the equation (66.3) taking into account the expression for K_{11}:

$$\mathbf{j}=e^2n\frac{\langle\tau\rangle}{m^*}\mathbf{E}-en\frac{\langle\tau\rangle}{m^*}\nabla F^*=\sigma\mathbf{E}-n\mu_d\nabla F^*. \tag{66.4}$$

For a non-degenerate semiconductor we may write from (65.3s)

$$\nabla F^*=kT\frac{\nabla n}{n}, \tag{66.5}$$

whence

$$\mathbf{j}=\sigma\mathbf{E}-kT\mu_d\nabla n=\mathbf{j}_E+\mathbf{j}_D, \tag{66.6}$$

where $\mathbf{j}_E=\sigma\mathbf{E}$ is the *ohmic*, or *drift*, *current* and

$$\mathbf{j}_D=-kT\mu_d\nabla n \tag{66.7}$$

is termed *diffusion current*. The reason for this term is that the current $\mathbf{j}_D$ results from carrier diffusion when the distribution in the crystal is non-uniform. The particle flow should be proportional to the concentration gradient and should be in the direction in which the concentration falls. Multiplying the particle flow by the charge and introducing a proportionality coefficient D we may write

$$\mathbf{j}_D=-eD\nabla n. \tag{66.8}$$

The quantity D is termed *diffusion coefficient*, its dimensionality being $[L^2T^{-1}]$ which follows from the definition (66.8). Comparing (66.8) with (66.7) we obtain the relation between the diffusion coefficient and the mobility of the charge carriers:

$$D=\frac{kT}{e}\mu_d, \tag{66.9}$$

which is valid both for holes and for electrons. The expression (66.9) is *the Einstein relation for mobility and diffusivity*. An analogous relation may be obtained for a degenerate semiconductor as well. Indeed, taking into account the relation between n and F^* we obtain

$$\frac{\nabla n}{n}=\frac{3}{2}\frac{\nabla F^*}{F^*};\quad \nabla F^*=\frac{2}{3}F^*\frac{\nabla n}{n}, \tag{66.10}$$

whence

$$D=\frac{2}{3}\frac{F^*}{e}\mu_d. \tag{66.11}$$

Substitute equation (66.6) for j into (66.2) to obtain

$$\frac{1}{e}\operatorname{div}(\sigma\mathbf{E}-eD\nabla n)+\frac{n-n_0}{\tau}=0, \tag{66.12}$$

or

$$n\mu_d\operatorname{div}\mathbf{E}+\mu_d(\mathbf{E}\nabla n)-D\nabla^2 n+\frac{n-n_0}{\tau}=0. \tag{66.13}$$

The expression (66.13) may be written both for holes and for electrons, the magnitude of the electric field in the equation (66.13) for electrons and holes being independent of the carrier charge sign. Since *electroneutrality is not affected by the current* we may, in compliance with the Poisson equation, write

$$\operatorname{div}\mathbf{E}=0, \tag{66.14}$$

and the equation (66.13) will assume the form

$$\nabla^2 n-\frac{\mu_d}{D}(\mathbf{E}\nabla n)-\frac{n-n_0}{D\tau}=0. \tag{66.15}$$

Since the non-equilibrium concentration is sustained by the current, ∇n should be co-linear with $\mathbf{E}$. Consider a unidimensional case setting $\mathbf{E}=(E, 0, 0)$ and write the equation (66.15) for the excess concentration $n-n_0$

$$\frac{d^2(n-n_0)}{dx^2}-\frac{\mu_d E}{D}\frac{d(n-n_0)}{dx}-\frac{n-n_0}{D\tau}=0. \tag{66.16}$$

Introduce the notation

$$D\tau=L^2,$$
$$\frac{\mu_d E}{D}=\frac{\mu_d E\tau}{D\tau}=\frac{l_E}{L^2}. \tag{66.17}$$

The equation (66.16) in new notation will be of the form

$$\frac{d^2(n-n_0)}{dx^2}-\frac{l_E}{L^2}\frac{d(n-n_0)}{dx}-\frac{n-n_0}{L^2}=0. \tag{66.18}$$

We shall look for the solution of the linear homogeneous equation in the form

$$n-n_0=Ae^{ax}. \tag{66.19}$$

Substituting (66.20) into (66.19) and cancelling out Ae^{ax} we obtain the characteristic equation

$$a^2-\frac{l_E a}{L^2}-\frac{1}{L^2}=0, \tag{66.20}$$

whose solution is

$$a=\frac{l_E}{2L^2}\pm\sqrt{\frac{l_E^2}{4L^4}+\frac{1}{L^2}}=\frac{1}{L}\left(\frac{l_E}{2L}\pm\sqrt{1+\frac{l_E^2}{4L^2}}\right). \qquad (66.21)$$

Introduce the quantity $l=-\frac{1}{a}$ and use it to write the solution (66.19):

$$n-n_0=Ae^{-x/l}. \qquad (66.22)$$

For the solution to be finite for $x\to\infty$ one should choose $a<0$ or $l>0$. A means excess concentration at the point $x=0$: $[n-n_0]|_{x=0}=A$. The condition $a<0$ requires that a minus sign be used in front of the radical in (66.21); therefore,

$$\frac{1}{l}=\frac{1}{L}\left(\sqrt{1+\frac{l_E^2}{4L^2}}-\frac{l_E}{2L}\right). \qquad (66.23)$$

We arrived at the conclusion that *the excess concentration decreases exponentially with the increase in x, the parameter of the exponent being termed drag distance l. l is equal to the distance at which the excess concentration decreases e times.*

Consider the case of no electric field: $E=0$. In this case $l_E = 0$ and

$$l=L=\sqrt{D\tau}. \qquad (66.24)$$

There is no electric field in the solid, and only the diffusion current flows in it:

$$\mathbf{j}=\mathbf{j}_D=-eD\nabla n. \qquad (66.25)$$

Since the sole cause of excess concentration variation is the non-equilibrium charge carrier diffusion, $l=L$ is termed *diffusion length.* For the unidimensional case

$$n(x)-n_0=\delta n(x)=\delta n(0)\,e^{-\frac{x}{L}} \qquad (66.26)$$

and

$$j=-eD\nabla_x n=\frac{eD\,\delta n(x)}{L}=e\frac{D\tau}{L\tau}\delta n=e\frac{L}{\tau}\delta n. \qquad (66.27)$$

Denote the ratio $\frac{L}{\tau}$ by v_D:

$$v_D=\frac{L}{\tau}=\frac{L^2}{L\tau}=\frac{D}{L}. \qquad (66.28)$$

The quantity v_D is termed diffusion velocity. Numerically it is equal to the velocity with which the non-equilibrium carriers cover the *diffusion length during their lifetime.* The diffusion velocity may be used to express the density of the diffusion current

$$j_{Dx} = ev_D(n - n_0) = j_{D0}e^{-\frac{x}{L}}. \tag{66.29}$$

The expression (66.29) shows that the diffusion current density decreases in accordance with the same law as the excess concentration, the cause of the latter being non-equilibrium carrier recombination. The expressions (66.24-29) are valid not only for $E=0$ but for $E \neq 0$ as well, provided $l_E \ll 2L$. Consider the opposite case—that of strong fields, $l_E^2 \gg 4L^2$. The meaning of this condition is obvious:

$$\frac{l_E^2}{4L^2} = \frac{(\mu_d E\tau)^2}{4L^2} = \frac{v_d^2\tau^2}{4L^2} = \frac{v_d^2}{4v_D^2} \gg 1. \tag{66.30}$$

The quantity l_E is numerically equal to the distance covered by non-equilibrium carriers travelling with drift velocity during their lifetime. Therefore, it is termed non-equilibrium carrier drift length. Hence, the field E should be regarded as strong if the drift length greatly exceeds the diffusion length. When determining the drag distance l one should, however, differentiate between the cases $l_E > 0$ and $l_E < 0$. Consider the case $l_E > 0$. From (66.23) we may write

$$\frac{1}{l} = \frac{1}{L}\left(\frac{l_E}{2L}\sqrt{1 + \frac{4L^2}{l_E^2}} - \frac{l_E}{2L}\right) = \frac{l_E}{2L^2}\left(1 + \frac{2L^2}{l_E^2} + \ldots - 1\right) \cong \frac{1}{l_E}, \tag{66.31}$$

i.e. $l = l_E$, the drag distance l is equal to the drift length:

$$l = l_E = \mu_d E\tau = v_d\tau \quad (l_E > 0). \tag{66.32}$$

Taking into account (66.22) and (66.32) write for the excess concentration $\delta n(x)$

$$\delta n(x) = \delta n(0)\, e^{-\frac{x}{l_E}} = \delta n(0)\, e^{-\frac{x}{\mu_d E\tau}}. \tag{66.33}$$

The meaning of this expression is as follows. An excess concentration established at the origin of co-ordinates $\delta n(0)$ is drawn by a strong electric field into the volume of the semiconductor to a distance of (2-3) l_E. If $\delta n(0) > 0$, the volume is flooded by non-equilibrium carriers, this effect being termed non-equilibrium carrier injection. Hole injection into the region $x > 0$ out of the point $x = 0$ takes place for $E > 0$, and electron injection for $E < 0$. In the case of a negative excess carrier concentration ($\delta n(0) < 0$), the volume will be depleted of charge carriers, this phenomenon being termed extraction. Consider the case $l_E < 0$ ($|l_E| > 0$).

From (66.23) we obtain for the drag distance

$$\frac{1}{l}=\frac{1}{L}\frac{|l_E|^2}{2L}=\frac{|E|}{L^2}; \quad l=\frac{L^2}{|l_E|}, \tag{66.34}$$

i.e. as the modulus of the field increases the drag distance tends to zero. The carrier concentration in the volume of the semiconductor as compared to the case E = 0 decreases for $\delta n(0)>0$ and increases for $\delta n(0)<0$, the former being termed exclusion, and the latter—accumulation of non-equilibrium charge carriers.

Consider an example of a uniform n-type extrinsic semiconductor sample with contacts at $x=0$ and $x=d$. If an external electric field $E>0$ is applied, holes (minority carriers) will be *injected* at the contact $x=0$ and *extracted* at the contact $x=d$. To compensate for the excess positive charge of the injected holes electrons will be *accumulated* near the contact $x=0$. They will be *excluded* from the region near $x=d$ because of hole deficiency in that region.

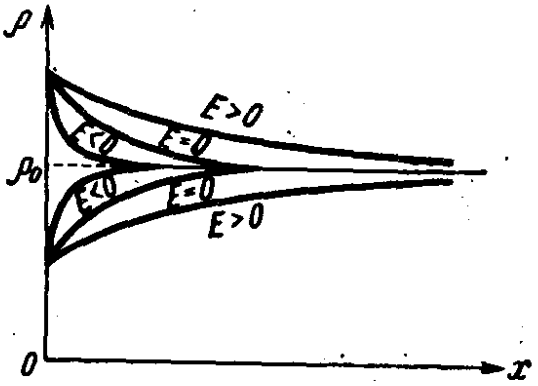


Fig. 98. The distribution of excess hole concentration in cases of injection, exclusion, extraction, and accumulation

Figure 98 shows the dependence of the distribution of $p-p_0$ on the direction of the field and on $\delta p(0)$. (It should be noted that in some papers the term exclusion is often applied to the phenomena we have termed extraction, and vice versa.)

When we considered the variations of charge carrier concentration we did not specify the type of the particles. Introducing the condition div $\mathbf{E}=0$ we, moreover, presumed the excess electron and hole concentrations to change in a similar way. To prove the validity of this approach consider a semiconductor in which a volume charge $\rho(\mathbf{r}, 0)$ has been set up by some means. This charge establishes an electric field $\mathbf{E}(\mathbf{r}, t)$ related to $\rho(\mathbf{r}, t)$ by the Poisson equation

$$\operatorname{div}\mathbf{E}=\frac{\rho}{\varepsilon\varepsilon_0}. \tag{66.35}$$

The electric field $\mathbf{E}$ causes a conductivity current $\mathbf{j}_E=\sigma\mathbf{E}$. It follows from the continuity equation

$$\frac{\partial\rho}{\partial t}=-\operatorname{div}\mathbf{j} \tag{66.36}$$

that

$$\frac{\partial\rho}{\partial t} = -\operatorname{div}\sigma\mathbf{E} = -\sigma\operatorname{div}\mathbf{E} = -\frac{\sigma\rho}{\varepsilon\varepsilon_0} = -\frac{\rho}{\tau^M}, \qquad (66.37)$$

$$\tau^M = \frac{\varepsilon\varepsilon_0}{\sigma}, \qquad (66.38)$$

τ^M having the dimensionality of time. The solution of the equation (66.37) is simple:

$$\rho(\mathbf{r},\ t) = \rho(\mathbf{r},\ 0)\,e^{-\frac{t}{\tau^M}}. \qquad (66.39)$$

Assess the value of τ^M setting $\varepsilon = 10$ and $\sigma = 10$ Sim·m^{-1} and taking into account that $\varepsilon_0 = \frac{10^7}{4\pi c^2} = 8.85416 \times 10^{-12}$ F/m. The result is $\tau^M \cong 10^{-11}$ s. *The quantity τ^M is termed Maxwell relaxation time* and is determined by the specific conductance of the substance. Thus, because of the *conductivity, a volume charge set up in a semiconductor exists, on the average, only during the time τ^M*. This leads to a result which is fundamental for the understanding of numerous problems of semiconductor- and semiconductor-device physics connected with non-equilibrium charge carriers. Namely, *should an excess concentration of majority carriers be set up in a semiconductor, the space charge and the excess conductivity would vanish, on the average, in the time τ^M. Should, however, an excess concentration of minority carriers be set up, Maxwell relaxation would result in their space charge being compensated by the majority carriers in the time τ^M and the non-equilibrium concentration of both the minority and the majority carriers would persist for a time equal to the non-equilibrium carrier lifetime τ_f*. Thus, the parts played by majority and minority carriers in the physical properties of the semiconductor are different in principle, and for this reason the injection or extraction of only *the minority carriers* is usually considered.

The initial space charge distribution remains constant in time as follows from the equation (66.39) because only the conductivity current is taken into account. In reality charge diffusion will take place as well; therefore, the distribution of the space charge in the crystal will be changed.

Summary of Sec. 66

1. Space charge compensation or dispersion due to conductivity is termed Maxwell relaxation. It is determined by the conductivity of the substance

in the SI system in the Gauss system

$$\tau^M = \frac{\varepsilon\varepsilon_0}{\sigma}; \qquad \tau^M = \frac{\varepsilon}{4\pi\sigma}. \tag{66.1s}$$

2. An excess concentration of majority carriers disperses in the time τ^M; an excess concentration of minority carriers vanishes together with the excess concentration of majority carriers produced by it in the time equal to the lifetime τ_f.

3. The variation of non-equilibrium minority carrier concentration in the absence of carrier generation at the expense of the external energy is described for stationary conditions by the equation

$$\frac{1}{e}\operatorname{div}\mathbf{j} + \frac{n-n_0}{\tau_f} = 0. \tag{66.2s}$$

4. In isothermal conditions, for $\mathbf{B}=0$, the current takes the form

$$\mathbf{j} = e^2K_{11}\mathbf{E} - eK_{11}\nabla F^* = \mathbf{j}_E + \mathbf{j}_D, \tag{66.3s}$$

where

$$\mathbf{j}_E = e^2K_{11}\mathbf{E} = en\mu_d\mathbf{E} \tag{66.4s}$$

is the conductivity current, and

$$\mathbf{j}_D = -n\mu_d\nabla F^* = -eK_{11}\nabla F^* = -eD\nabla n \tag{66.5s}$$

is the diffusion current.

5. The diffusion coefficient and the mobility of charge carriers are related by means of the Einstein relation

$$D_n = \frac{kT}{e_n}\mu_{dn}, \quad D_p = \frac{kT}{e_p}\mu_{dp} \tag{66.6s}$$

$$D_n = \frac{2F_n^*}{3e_n}\mu_{dn}, \quad D_p = \frac{2F_p^*}{3e_p}\mu_{dp} \tag{66.7s}$$

for the non-degenerate and the degenerate semiconductor, respectively.

6. In the unidimensional case the excess concentration of minority carriers is determined by the equation

$$\delta n(x) = \delta n(0)\, e^{-\frac{x}{l}}, \tag{66.8s}$$

where l is termed drag distance. l is related to the diffusion length $L = \sqrt{D\tau_f}$ and to the drift length $l_E = \mu_d E\tau_f$ by means of the expression (66.23). In weak electric fields the minority carrier distribution is determined by diffusion; in strong fields—by drift.

67. SURFACE RECOMBINATION

The surface of a semiconductor contains a number of defects which act as recombination centres. This results in surface recombination diminishing non-equilibrium charge carrier concentration in layers adjoining the surface. Experimentally determined lifetimes of the same semiconductor sample will be different for different surface treatments such as grinding, polishing, etching, plating, i.e. the lifetime depends on surface treatment. Surface recombination may be observed visually or with the aid of an image converter. Figure 99 shows the photograph of a gallium arsenide diode operating in conditions of carrier injection by high current pulses. It may be seen from Fig. 99 that the layers near the surface radiate light while the volume remains dark (except the *p-n* junction region where stimulated recombination takes place). *Surface recombination is usually described by the so-called surface recombination velocity which may be introduced in the following way.*

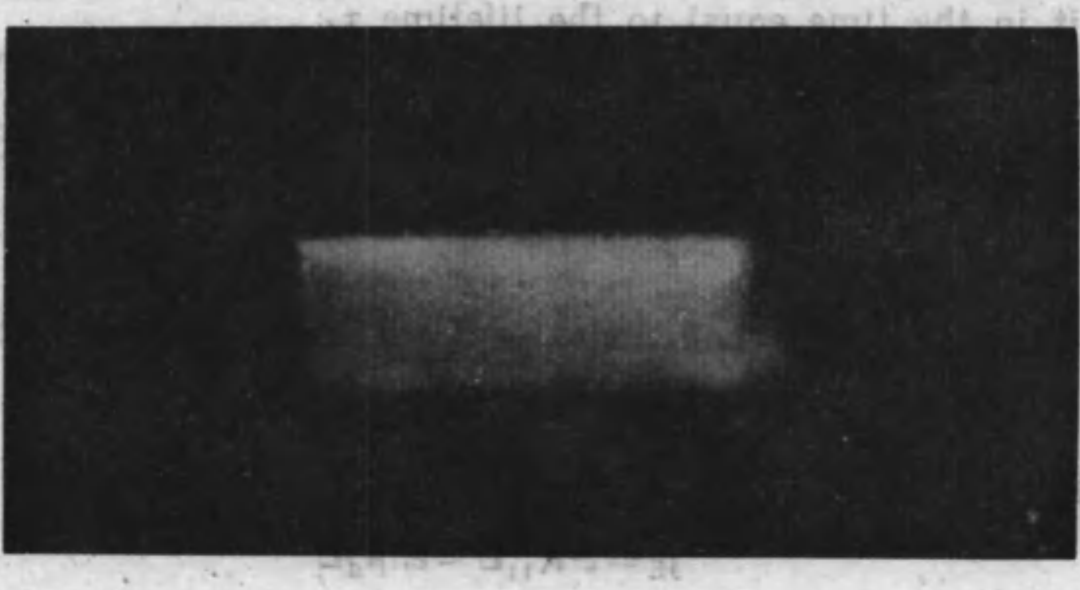


Fig. 99. The luminescence of a gallium arsenide diode surface due to surface recombination

Write the continuity equation in the form

$$\frac{\partial n}{\partial t} = -\frac{1}{e}\operatorname{div}\mathbf{j} - \frac{n-n_0}{\tau_f} + G. \tag{67.1}$$

This equation was already solved by us in the unidimensional case in the assumptions of infinite y and z dimensions of the semiconductor sample and of $G=0$. Take account of the fact that now the sample is limited. The external electric field directed along the x-axis of the sample provides for minority carrier injection. In order to simplify the equation (67.1) we will assume the electric field to be absent, $\mathbf{E}=0$, and the excess concentration to be produced by a source with a carrier generation rate G independent of the coordinate. The electric field in this case will everywhere be zero. Should we consider a sample in the form of a plate of infinite y and x dimensions, the continuity equation would again become unidimensional. Because of surface recombination the excess carrier concentration near the surface of the sample should be less than in its volume away from the surface. The non-equilibrium carrier concentration gradient should result in a diffusion flow directed from the volume of the semiconductor to its surface. Let the recombination current at the surface be j_z^s. The continuity equation (67.1) for stationary conditions may be written in the form

$$0 = -\frac{1}{e}\operatorname{div}\mathbf{j} - \frac{n-n_0}{\tau_f} + \mathrm{G}. \tag{67.2}$$

Substituting $-eD\nabla_z n$ for $\mathbf{j}$ we obtain

$$D\frac{d^2n}{dz^2} - \frac{n-n_0}{\tau_f} + \mathrm{G} = 0, \tag{67.3}$$

or

$$\frac{d^2(n-n_0)}{dz^2} - \frac{n-n_0}{L^2} + \frac{\mathrm{G}}{D} = 0. \tag{67.4}$$

The general solution of the homogeneous equation is

$$\delta n(z) = n(z) - n_0 = Ae^{-\frac{z}{L}} + Be^{\frac{z}{L}}. \tag{67.5}$$

The partial solution of the inhomogeneous equation (67.3) may be chosen as follows:

$$\delta n = \mathrm{G}\tau_f. \tag{67.6}$$

The general solution of the inhomogeneous equation is equal to their sum:

$$\delta n(z) = \mathrm{G}\tau_f\left[1 - A'e^{-\frac{z}{L}} - B'e^{\frac{z}{L}}\right], \tag{67.7}$$

where

$$A' = -\frac{A}{\mathrm{G}\tau_f}, \quad B' = -\frac{B}{\mathrm{G}\tau_f}. \tag{67.8}$$

The quantities A' and B' may be found from the boundary conditions

$$\begin{gathered}\delta n(0) = \mathrm{G}\tau_f[1 - A' - B'], \\ \delta n(d) = \mathrm{G}\tau_f\left[1 - A'e^{-\frac{d}{L}} - B'e^{\frac{d}{L}}\right],\end{gathered} \tag{67.9}$$

where d is the thickness of the plate. Express the boundary conditions in terms of the current $j_z^s(0)$ and $j_z^s(d)$. Since

$$j_z^s = -eD\frac{dn}{dz} = -eDG\tau_f\left[\frac{A'}{L}e^{-\frac{z}{L}} - \frac{B'}{L}e^{\frac{z}{L}}\right], \qquad (67.10)$$

it follows that

$$j_z^s(0) = -eGL(A' - B'), \qquad (67.11)$$

$$j_z^s(d) = -eGL\left(A'e^{-\frac{d}{L}} - B'e^{\frac{d}{L}}\right). \qquad (67.12)$$

The expressions for A' and B' may be obtained from (67.11) and (67.12)

$$A' = \frac{-j_z^s(0)\,e^{\frac{d}{L}} + j_z^s(d)}{2eLG\sinh\frac{d}{L}}, \qquad (67.13)$$

$$B' = \frac{-j_z^s(0)\,e^{-\frac{d}{L}} + j_z^s(d)}{2eLG\sinh\frac{d}{L}}. \qquad (67.14)$$

The surface recombination current may be expressed in terms of excess carrier concentration on the surface, δn^s:

$$j^s = e\delta n^s s \qquad (67.15)$$

where s is the proportionality coefficient connecting the current j^s with the excess concentration. Evidently, the dimensionality of s is that of velocity: $[s] = [LT^{-1}]$. The quantity s became known as surface recombination velocity. The coefficients A' and B' may be expressed in terms of recombination velocities s_1 $(z=0)$ and s_2 $(z=d)$. To this end the expressions (67.9-11) and (67.15) should be used for δn:

$$\delta n(0) = G\tau_f[1 - A' - B'] = \frac{j^s(0)}{es_1} = -\frac{GL}{s_1}[A' - B'], \qquad (67.16)$$

$$\delta n(d) = G\tau_f\left[1 - A'e^{-\frac{d}{L}} - B'e^{\frac{d}{L}}\right] = \frac{j^s(d)}{es_2} = -\frac{GL}{s_2}\left[A'e^{-\frac{d}{L}} - B'e^{\frac{d}{L}}\right]. \qquad (67.17)$$

Solving the system of equations (67.16) and (67.17) we obtain the expressions for A' and B':

$$A' = \frac{s_1(v_D+s_2)\,e^{\frac{d}{L}} - s_2(v_D+s_1)}{2(s_1s_2 - v_D^2)\sinh\frac{d}{L} + 2v_D(s_1-s_2)\cosh\frac{d}{L}}, \qquad (67.18)$$

$$B' = \frac{s_1(v_D-s_2)\,e^{-\frac{d}{L}} - s_2(v_D-s_1)}{2(s_1s_2 - v_D^2)\sinh\frac{d}{L} + 2v_D(s_1-s_2)\cosh\frac{d}{L}}. \qquad (67.19)$$

$v_D = \frac{L}{\tau_f}$ in expressions (67.18-19) is the diffusion velocity. The sigrs of s_1 and s_2 are opposite. Since the current directed to the $(z=d)$-plane is positive, and to the $(z=0)$-plane—negative, we will assume $s_1 < 0$ and $s_2 > 0$.

Consider the distribution of $\delta n(z)$ for different values of s_1 and s_2.

1. **There is no surface recombination,** $s_1 = s_2 = 0$. It follows from (67.18, 19) that $A' = B'$ and $\delta n(z) = G\tau_f$, i.e. δn is independent of the co-ordinate.

2. **The surface recombination velocity is infinite,** $|s_1| = s_2 = \infty$. Physically, this means that the concentration of excess carriers reaching the surface turns zero because of the recombination. It follows from (67.18, 19) that

$$A' = \frac{e^{\frac{d}{L}} - 1}{2\sinh\frac{d}{L}}; \quad B' = \frac{1 - e^{-\frac{d}{L}}}{2\sinh\frac{d}{L}} \qquad (67.20)$$

and

$$\delta n(z) = G\tau_f\left[1 - \frac{\sinh\frac{z}{L} + \sinh\frac{d-z}{L}}{\sinh\frac{d}{L}}\right]. \qquad (67.21)$$

Also (67.21) yields $\delta n(0) = \delta n(d) = 0$.

3. In the same way other cases may be analysed. For instance, that of $s_1 = 0$ and $s_2 = \infty$

$$A' = B' = \frac{1}{2\cosh\frac{d}{L}}, \qquad (67.22)$$

and

$$\delta n(z) = G\tau_f\left[1 - \frac{\cosh\frac{z'}{L}}{\cosh\frac{d}{L}}\right]. \qquad (67.23)$$

The excess carrier concentration turns zero for $z=d$: $\delta n(d)=0$, since $s_2=\infty$. At the point $z=0$ $\delta n(0)=G\tau_f\left[1-\frac{1}{\cosh\frac{d}{L}}\right]$. If the diffusion length is small and the thickness of the sample is great, $\delta n(0)=G\tau_f$. Should, however, d be comparable to L, $\delta n<G\tau_f$ since a great recombination velocity on the surface ($z=d$) results

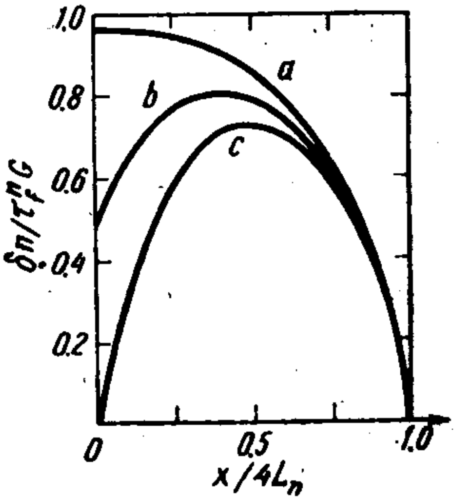


Fig. 100. The dependence of excess carrier concentration on surface recombination velocity:

(a) $s_1=0$; $|s_1|=v_D$; (c) $|s_1|=\infty$; $s_2=\infty$. In all cases the sample thickness is $4L_n$

in a marked carrier depletion in the semiconductor volume. Generally, the depletion should be noticeable at distances from the surface of the order of the diffusion length. Figure 100 shows the distribution of $\delta n(z)$ for the case $d=4L$ for different values of s_1 and for $s_2=\infty$.

The surface recombination velocities observed in germanium were from 50 to over 10^6 cm/s.

It should be remarked that the current caused by surface recombination cannot result in electric fields since the condition of electroneutrality is not violated in this case.

Sometimes the concept of effective lifetime τ_f^{eff} is introduced to describe the part played by surface recombination. τ_f^{eff} is determined by the condition

$$\frac{1}{\tau_f^{eff}}=\frac{1}{\tau_f^v}+\frac{1}{\tau_f^s}, \qquad (67.24)$$

where τ_f^v is the lifetime in the sample due only to volume recombination ($s=0$; it is equal to the lifetime in the volume τ_f), and τ_f^s is the life in the sample due only to surface recombination ($\tau_f\to\infty$). For the same values of τ_f and s τ_f^{eff} will depend on the geometry of the sample.

Chapter VII

CONTACT PHENOMENA IN SEMICONDUCTORS

68. DEBYE LENGTH

When two substances are brought into contact, they start to exchange charge carriers and this results in variations of semiconductor properties not only on the contact surface but in volume as well.

The practical application of semiconductors and the study of many of their properties necessarily involves the connection of the semiconductor sample into a circuit made up of various materials. In order to understand numerous contact phenomena the variation of the properties of a semiconductor in an electric field should be considered.

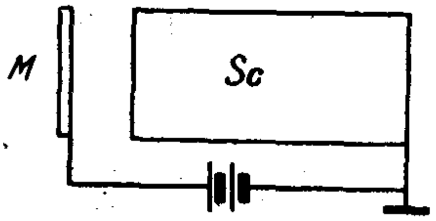


Fig. 101. A semiconductor sample in a uniform electric field

Consider a semiconductor sample placed into the electric field of a capacitor assumed, for the sake of simplicity, to be uniform (Fig. 101). The field will affect charge redistribution in the semiconductor with the result that a space charge $\rho(\mathbf{r})$ and an electric field $\mathbf{E}$ connected with the space charge by the Poisson equation

$$\operatorname{div} \varepsilon\varepsilon_0 \mathbf{E}(\mathbf{r}) = \rho(\mathbf{r}) \tag{68.1}$$

will be established.

Should the field $\mathbf{E}$ be expressed through the potential the Poisson equation (68.1) would assume the form

$$\nabla^2 \varphi(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\varepsilon\varepsilon_0}. \tag{68.2}$$

The potential energy of the electron $V(\mathbf{r})$ is related to $\varphi(\mathbf{r})$ as follows: $V(\mathbf{r}) = e\varphi(\mathbf{r})$. It was, however, demonstrated in Sec. 19 that the potential electric field $V(\mathbf{r})$ bends the energy bands so that $E(\mathbf{r}) = E^{\prime}+V(\mathbf{r})$ or

$$E_c(\mathbf{r}) = E_c + V(\mathbf{r}). \tag{68.3}$$

All the energy levels including the discrete levels of the forbidden band are displaced. But if the semiconductor is in the state of thermal equilibrium, the position of its Fermi level is everywhere the same. Therefore, the distance between it and the energy bands changes:

in the absence of a field

$$E_c - F; \quad F - E_v; \tag{68.4}$$

in the presence of a field

$$E_c + V(\mathbf{r}) - F; \quad F - [E_v + V(\mathbf{r})]. \tag{68.5}$$

Comparing (68.5) with (68.4) we see that if the distance between F and E_c increases by the amount $V(\mathbf{r})$, the distance between F and E_v decreases by the same amount, and vice versa. The variation of the distance between F and the energy level results in the variation of carrier distribution over energy levels.

According to (30.5) the electroneutrality equation is

$$n + n_d - p - p_a = N_d - N_a. \tag{68.6}$$

The field $V(\mathbf{r})$ affects different terms of (68.6) in a different way, creating a space charge

$$\rho(\mathbf{r}) = e(\delta n + \delta n_d - \delta p - \delta p_a)(e = e_u). \tag{68.7}$$

Denoting equilibrium particle concentrations by the index 0 and leaving the corresponding values in the presence of a field without the index 0, we obtain

$$\begin{aligned} \delta n &= n(\mathbf{r}) - n_0; \qquad \delta p = p(\mathbf{r}) - p_0; \\ \delta n_d &= n_d(\mathbf{r}) - n_{d0}; \quad \delta p_a = p_a(\mathbf{r}) - p_{a0}. \end{aligned} \tag{68.8}$$

For a non-degenerate semiconductor

$$n(\mathbf{r}) = N_c e^{-\frac{E_c(\mathbf{r}) - F}{kT}} = N_c e^{-\frac{E_c - F}{kT}} \cdot e^{-\frac{V(\mathbf{r})}{kT}} = n_0 e^{-\frac{V(\mathbf{r})}{kT}} \tag{68.9}$$

$$p(\mathbf{r}) = p_0 e^{\frac{V(\mathbf{r})}{kT}}, \tag{68.10}$$

$$n_d(\mathbf{r}) = \frac{N_d}{\frac{1}{2} e^{\frac{E_d + V(\mathbf{r}) - F}{kT}} + 1};$$

$$p_a(\mathbf{r}) = \frac{N_a}{\frac{1}{2} e^{\frac{F - E_a - V(\mathbf{r})}{kT}} + 1}. \tag{68.11}$$

Multiplying (68.2) by the electron charge $e=e_n$ and taking into account (68.7-11) we obtain the equation for $V(\mathbf{r})$

$$\nabla^2 V(\mathbf{r}) = -\frac{e\rho(\mathbf{r})}{\varepsilon\varepsilon_0} = -\frac{e^2}{\varepsilon\varepsilon_0}(\delta n+\delta n_d-\delta p-\delta p_a) =$$

$$= -\frac{e^2}{\varepsilon\varepsilon_0}\left\{\left[n_0\left(e^{-\frac{V(\mathbf{r})}{kT}}-1\right)-p_0\left(e^{\frac{V(\mathbf{r})}{kT}}-1\right)\right]+(\delta n_d-\delta p_a)\right\}. \quad (68.12)$$

The expressions for δn_d and δp_a are quite complicated, therefore we shall retain the concise notation.

The first integral of the equation (68.12) may be calculated with the aid of a standard method. Multiply (68.12) by $\nabla V(\mathbf{r})$, and the left-hand member of the equation will assume the form

$$[\nabla V(\mathbf{r})]\,\nabla^2 V(\mathbf{r}) = \frac{1}{2}\nabla[\nabla V(\mathbf{r})]^2. \quad (68.13)$$

The equations in square brackets of (68.12) may be transformed as follows:

$$-\frac{e^2}{\varepsilon\varepsilon_0}\left[n_0\left(e^{-\frac{V(\mathbf{r})}{kT}}-1\right)-p_0\left(e^{\frac{V(\mathbf{r})}{kT}}-1\right)\right]\nabla V(\mathbf{r}) =$$

$$= \nabla\left\{-\frac{e^2}{\varepsilon\varepsilon_0}\left[n_0\left(-e^{-\frac{V(\mathbf{r})}{kT}}kT-V(\mathbf{r})\right)-p_0\left(e^{\frac{V(\mathbf{r})}{kT}}kT-V(\mathbf{r})\right)\right]\right\}. \quad (68.14)$$

Suppose that in the same way some function $\Omega(\mathbf{r})$ may be chosen, so that

$$(\delta n_d-\delta p_a)\,\nabla V(\mathbf{r}) = \nabla\Omega(\mathbf{r}).. \quad (68.15)$$

In this case we will be able to write the first integral in the equation (68.12) in the form

$$\pm|\nabla V(\mathbf{r})| = \left\{\frac{2e^2kT}{\varepsilon\varepsilon_0}\left[n_0\left(e^{-\frac{V(\mathbf{r})}{kT}}-\frac{V(\mathbf{r})}{kT}\right)+p_0\left(e^{\frac{V(\mathbf{r})}{kT}}-\frac{V(\mathbf{r})}{kT}\right)\right]-\right.$$

$$\left.-\frac{2e^2}{\varepsilon\varepsilon_0}\Omega(\mathbf{r})+C\right\}^{1/2}. \quad (68.16)$$

The integration constant C may be found from the boundary conditions for V and ∇V. Further solution of the equation (68.16) in a generalized form is practically impossible, and for this reason we shall now consider some specific cases.

1. **Intrinsic semiconductor.** In an intrinsic semiconductor $N_d = N_a = 0$. Therefore $\delta n_d = \delta p_a = 0$. Moreover, $n_0 = p_0 = n_i$, and the equation (68.16) assumes the form

$$\pm|\nabla V(\mathbf{r})| = \sqrt{\frac{2e^2 n_i kT}{\varepsilon\varepsilon_0}}\left[e^{\frac{V(\mathbf{r})}{kT}}+e^{-\frac{V(\mathbf{r})}{kT}}+C\right]^{1/2}. \quad (68.17)$$

Should we take as the origin of V the potential energy at the points of the sample where the electric field intensity is zero, from the condition $\nabla V = V = 0$ we would obtain

$$0 = [2 + C]^{1/2}; \; C = -2 \tag{68.18}$$

and

$$\pm |\nabla V(\mathbf{r})| = 2\sqrt{\frac{e^2 n_i kT}{\varepsilon\varepsilon_0}}\left(\cosh\frac{V(\mathbf{r})}{kT} - 1\right)^{1/2} =$$

$$= 2\sqrt{\frac{2e^2 n_i kT}{\varepsilon\varepsilon_0}}\sinh\frac{V(\mathbf{r})}{2kT}. \tag{68.19}$$

We shall confine ourselves to the unidimensional case

$$\frac{dV(x)}{dx} = \pm 2\sqrt{\frac{2e^2 n_i kT}{\varepsilon\varepsilon_0}}\sinh\frac{V(x)}{2kT}. \tag{68.20}$$

Separating the variables in (68.20) we obtain

$$\pm x + C_1 = \int \frac{1}{2}\sqrt{\frac{\varepsilon\varepsilon_0}{2e^2 n_i kT}}\frac{dV}{\sinh\frac{V}{kT}}. \tag{68.21}$$

The integral in (68.21) may be calculated as follows:

$$\int\frac{d\xi}{\sinh\xi} = \int\frac{d\xi}{2\sinh\frac{\xi}{2}\cosh\frac{\xi}{2}} = \int\frac{d\frac{\xi}{2}}{\frac{\sinh\frac{\xi}{2}}{\cosh\frac{\xi}{2}}\cosh^2\frac{\xi}{2}} =$$

$$= \int\frac{d\tanh\frac{\xi}{2}}{\tanh\frac{\xi}{2}} = \ln\tanh\frac{\xi}{2}. \tag{68.22}$$

Therefore we obtain for (68.21)

$$\pm x + C_1 = \sqrt{\frac{\varepsilon\varepsilon_0 kT}{2e^2 n_i}}\ln\tanh\frac{V(x)}{4kT}. \tag{68.23}$$

Introduce the notation

$$\sqrt{\frac{\varepsilon\varepsilon_0 kT}{2e^2 n_i}} = L^D, \tag{68.24}$$

L^D is termed Debye screening length. Since the value of the hyperbolic tangent cannot exceed unity, and $x > 0$, only the term with the minus sign in front of x should be retained:

$$-x + C_1 = L^D \ln\tanh\frac{V(x)}{4kT}. \tag{68.25}$$

The quantity C_1 may be determined from the condition on the surface: $V(0)=V_s$ for $x=0$. Therefore

$$C_1 = L^D \ln\tanh\frac{V_s}{4kT} \qquad (C_1<0). \tag{68.26}$$

Express $\tanh\frac{V(x)}{4kT}$ in terms of x taking into account (68.25):

$$\tanh\frac{V(x)}{4kT} = e^{\frac{C_1-x}{L^D}}. \tag{68.27}$$

Solving the equation (68.27) for $V(x)$ we obtain

$$V(x) = 2kT\ln\left(\frac{1+e^{\frac{C_1-x}{L^D}}}{1-e^{\frac{C_1-x}{L^D}}}\right) = 2kT\ln\left[\frac{1+e^{-\frac{x}{L^D}}\tanh\frac{V_s}{4kT}}{1-e^{-\frac{x}{L^D}}\tanh\frac{V_s}{4kT}}\right]. \tag{68.28}$$

The equation (68.28) describes band displacement as a function of the co-ordinate x and V_s. It may be seen that the displacement changes monotonically from V_s for $x=0$ to zero for $x\longrightarrow\infty$. Since the hyperbolic tangent cannot exceed unity we may simplify the expression for $V(x)$ for small and large x. For $x\cong(2\text{-}3)L^D$ the exponent is much less than unity and we may use the relation $\ln(1+\xi)\cong\xi$ to write

$$V(x)\cong 2kT\cdot 2\tanh\frac{V_s}{4kT}e^{-\frac{x}{L^D}} = 4kT\tanh\frac{V_s}{4kT}e^{-\frac{x}{L^D}}. \tag{68.29}$$

If $\frac{V_s}{4kT}\ll 1$, the expression (68.29) will be valid for any x:

$$V(x) = V_s e^{-\frac{x}{L^D}}. \tag{68.30}$$

Consider the case of small x. In order to expand the expression (68.28) into a series in x at the point $x=0$ find $\frac{dV}{dx}$ and $\frac{d^2V}{dx^2}$. Leaving out simple but lengthy computations we present the result:

$$\frac{dV(x)}{dx} = -\frac{4kT}{L^D}\tanh\frac{V_s}{4kT}\frac{e^{-\frac{x}{L^D}}}{1-\tanh^2\frac{V_s}{4kT}e^{-\frac{2x}{L^D}}}; \tag{68.31}$$

$$\frac{d^2V(x)}{dx^2} = \frac{4kT\tanh\frac{V_s}{4kT}}{L^{D^2}}\,\frac{\left(1+\tanh^2\frac{V_s}{4kT}e^{-\frac{2x}{L^D}}\right)e^{-\frac{x}{L^D}}}{\left(1-\tanh^2\frac{V_s}{4kT}e^{-\frac{2x}{L^D}}\right)^2}. \tag{68.32}$$

Setting $x=0$ we obtain

$$\left.\frac{dV(x)}{dx}\right|_{x=0}=-e\mathrm{E}_s=-\frac{4kT}{L^D}\tanh\frac{V_s}{4kT}\frac{1}{1-\tanh^2\frac{V_s}{4kT}}=$$

$$=-\frac{2kT}{L^D}\sinh\frac{V_s}{2kT}. \qquad (68.33)$$

Confining ourselves to the case of small x we write

$$V(x)=V_s-\frac{2kT}{L^D}\sinh\frac{V_s}{2kT}x=V_s\left(1-\frac{2kT}{V_s}\sinh\frac{V_s}{2kT}\frac{x}{L^D}\right). \qquad (68.34).$$

Thus, for $x \ll L^D$ the potential energy $V(x)$ varies linearly with x, the variation becoming exponential for $x > L^D$. The variation of $V(x)$ for large $V_s (V_s \gg kT)$ takes place in a region the thickness of which is of the order of L^D.

The intensity of the electric field $\mathrm{E}(x)$ is determined from the relation (68.31) as follows:

$$\mathrm{E}(x)=-\frac{1}{e}\frac{dV(x)}{dx}=\frac{4kT}{eL^D}\tanh\frac{V_s}{4kT}\frac{e^{-\frac{x}{L^D}}}{1-\tanh^2\frac{V_s}{4kT}\cdot e^{-\frac{2x}{L^D}}}. \qquad (68.35)$$

On the surface $x=0$ E assumes, according to (68.33), the value

$$\mathrm{E}(0)=\mathrm{E}_s=\frac{2kT}{eL^D}\sinh\frac{V_s}{2kT}. \qquad (68.36)$$

To determine the space charge density $\rho(x)$ use (68.2) and (68.32):

$$\rho(x)=-\frac{\varepsilon\varepsilon_0}{e}\frac{d^2V(x)}{dx^2}=$$

$$=-\frac{4\varepsilon\varepsilon_0 kT\tanh\frac{V_s}{4kT}\left[1+\tanh^2\frac{V_s}{4kT}e^{-\frac{2x}{L^D}}\right]e^{-\frac{x}{L^D}}}{eL^{D2}\left[1-\tanh^2\frac{V_s}{4kT}e^{-\frac{2x}{L^D}}\right]^2}. \qquad (68.37)$$

Calculate the space charge density on the surface of the semiconductor, ρ_s:

$$\rho_s=\rho(0)=-\frac{4\varepsilon\varepsilon_0 kT\tanh\frac{V_s}{4kT}\left(1+\tanh^2\frac{V_s}{4kT}\right)}{eL^{D2}\left(1-\tanh^2\frac{V_s}{4kT}\right)^2}. \qquad (68.38)$$

Taking into account that

$$\frac{\tanh \xi\,(1+\tanh^2 \xi)}{(1-\tanh^2 \xi)^2}=\frac{1}{4}\sinh 4\xi, \tag{68.39}$$

and substituting the expression (68.24) for L^D, we obtain

$$\rho_s=-2en_i \sinh \frac{V_s}{kT}=en_i\left(e^{-\frac{V_s}{kT}}-e^{\frac{V_s}{kT}}\right). \tag{68.40}$$

The expression (68.40) may be obtained directly from (68.7) with the aid of (68.9-10). The sign of space charge density on the surface ρ_s depends on the sign of V_s. For $V_s>0$ (the energy bands are displaced upwards) $\rho_s>0$; for $V_s<0$ $\rho_s<0$. The magnitude $|\rho_s|$ is independent of the V_s sign.

The result obtained means that a space charge is established at the semiconductor surface due to an increase in electron ($V_s<0$) or hole ($V_s>0$) concentration. The space-charge layer is not thick since for small x the $V(x)$ function is proportional to x, and for this reason $\rho(x)$ decreases exponentially already for small x, the rate of decrease being higher for larger x. Since the concentration of charge carriers at the surface may greatly exceed their concentration in the volume

$$n_s \cong n_i e^{\frac{V_s}{kT}}, \tag{68.41}$$

the conductivity of the surface layers may experience a sharp rise. This phenomenon is known as the *field effect*. External field is not the only possible cause of the increase in carrier concentration in the surface layers. As was mentioned in Sec. 22, the semiconductor surface may accommodate a substantial number of surface states whose energy levels may be inside the forbidden band. The surface states are capable of trapping free charge carriers—electrons or holes—and thus creating a *surface charge* on the semiconductor surface. This surface charge will set up an electric field which will affect the volume properties of the semiconductor in the same way as an external field. Figure 102 shows the dependence of V, E and ρ on x.

2. **Extrinsic semiconductor.** Consider a semiconductor doped, for example, with a donor impurity in the temperature range in which the impurity is completely ionized:

$$n_0=N_d^+=N_d, \quad p_0=\frac{n_i^2}{N_d}. \tag{68.42}$$

Neglecting the variation of the impurity ionization coefficient we obtain, in accordance with (68.16),

$$\frac{dV(x)}{dx}=\left\{\frac{2e^2kTn_0}{\varepsilon\varepsilon_0}\left[e^{-\frac{V}{kT}}+\frac{V}{kT}+\frac{n_i^2}{n_0^2}\left(e^{\frac{V}{kT}}-\frac{V}{kT}\right)+C\right]\right\}^{1/2}. \tag{68.43}$$

Determining the integration constant from the same condition $V=0$ for $\nabla_x V=0$ we obtain

$$1+\frac{n_i^2}{n_0^2}+C=0;\quad C=-\left(1+\frac{n_i^2}{n_0^2}\right) \tag{68.44}$$

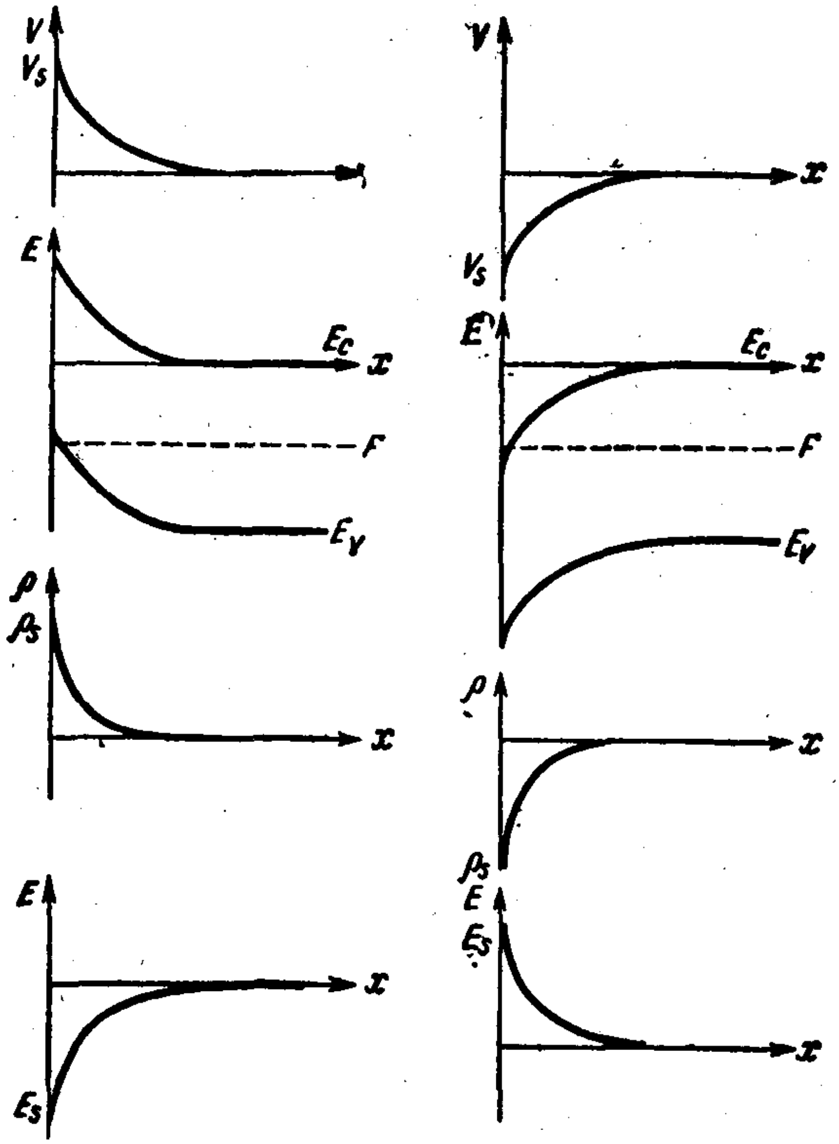


Fig. 102. The bending of the energy bands in an external field, the space charge, and the electric field in a semiconductor

and

$$\frac{dV(x)}{dx}=\pm\left\{\frac{2e^2n_0kT}{\varepsilon\varepsilon_0}\left[e^{-\frac{V}{kT}}+\frac{V}{kT}\left(1-\frac{n_i^2}{N_d^2}\right)+\right.\right.$$
$$\left.\left.+\frac{n_i^2}{N_d^2}e^{\frac{V}{kT}}-\left(1+\frac{n_i^2}{N_d^2}\right)\right]\right\}^{1/2} \tag{68.45}$$

If the doping level is sufficiently high, $n_i^2 \ll N_d^2$. For instance, for germanium with $N_d \doteq 10^{16}$ at $T \cong 300$ K $\frac{n_i^2}{N_d^2} \cong 10^{-6}$. Neglecting

$\frac{n_i^2}{N_d^2}$ as compared to unity we obtain for (68.45)

$$\frac{dV(x)}{dx} = \pm \left\{ \frac{2e^2kTn_0}{\varepsilon\varepsilon_0} \left[e^{-\frac{V(x)}{kT}} + \frac{V}{kT} - 1 + \frac{n_i^2}{N_d^2} e^{\frac{V}{kT}} \right] \right\}^{1/2} \qquad (68.46)$$

For $V \ll kT$ the equation (68.46) may be simplified still further:

$$\frac{dV(x)}{dx} = \pm \left\{ \frac{2e^2kTn_0}{\varepsilon\varepsilon_0} \frac{n_i^2}{N_d^2} \right\}^{1/2} e^{\frac{V}{2kT}}, \quad \left(\frac{n_i^2}{N_d^2} \gg \frac{V}{kT} \right) \qquad (68.47)$$

and

$$\frac{dV(x)}{dx} = \pm \left\{ \frac{e^2n_0}{\varepsilon\varepsilon_0 kT} \right\}^{1/2} V, \quad \left(\frac{n_i^2}{N_d^2} \ll \frac{V}{kT} \right) \qquad (68.48)$$

Introducing the notation

$$\sqrt{\frac{\varepsilon\varepsilon_0 kT}{e^2n_0}} = L^D, \qquad (68.49)$$

we write

$$\frac{dV(x)}{dx} = \pm \frac{V}{L^D}, \qquad (68.50)$$

whence the solution which remains finite for $x \to \infty$ is

$$V(x) = V_s e^{-\frac{x}{L^D}} \qquad (68.51)$$

The bending of the energy bands and the variation of electron and hole concentrations take place in extrinsic as well as in intrinsic semiconductors. $V_s < 0$ results in an increase in electron concentration, and $V_s > 0$—in an increase in hole concentration. For this reason *the surface layer is depleted of the majority carriers and is flooded with the minority carriers.*

In the case described by the equation (68.47) the semiconductor behaves likewise.

The equation (68.46) for $|V| \gg kT$ may be simplified as follows. For $V < 0$ we retain the term of maximum value $e^{\frac{V}{kT}}$ and obtain

$$\frac{dV(x)}{dx} = \pm \left(\frac{2e^2kTn_0}{\varepsilon\varepsilon_0} \right)^{1/2} e^{-\frac{V}{2kT}}. \qquad (68.52)$$

The solution of the equation (68.52) is

$$e^{\frac{V}{2kT}} = \pm \left(\frac{e^2n_0}{2\varepsilon\varepsilon_0 kT} \right)^{1/2} x + C. \qquad (68.53)$$

Putting $x = 0$; $V(0) = V_s$, we obtain

$$C = e^{\frac{V_s}{2kT}} \quad (C \ll 1). \tag{68.54}$$

Expression (68.53) gives us the asymptotic solution

$$V(x) = 2kT \ln\left(C + \frac{x}{L^{D'}}\right); \quad L^{D'} = \sqrt{\frac{2\varepsilon\varepsilon_0 kT}{e^2 n_0}}, \tag{68.55}$$

which is valid for x for which $-\frac{V(x)}{kT} \gg 1$. Find the variation of the electron and hole concentrations:

$$n(x) = n_0 e^{-\frac{V}{kT}} = \frac{n_0}{\left(C + \frac{x}{L^D}\right)^2}, \tag{68.56}$$

i.e. the electron concentration decreases as the square of a hyperbolic function, the surface layer being flooded with electrons:

$$n_s = n(0) = n_0 e^{-\frac{V_s}{kT}} \quad (V_s < 0). \tag{68.57}$$

For $V > 0$ we retain the term $\frac{V}{kT}$ in the equation (68.46),

$$\frac{dV(x)}{dx} = \pm\left(\frac{2e^2 n_0}{\varepsilon\varepsilon_0}\right)^{1/2} V^{1/2}, \tag{68.58}$$

and

$$V^{1/2}(x) = \pm 2\left(\frac{2e^2 n_0}{\varepsilon\varepsilon_0}\right)^{1/2} x + C. \tag{68.59}$$

The integration constant may be found from the boundary conditions $V(0) = V_s = C^2$, i.e., $C = V_s^{1/2}$. Therefore we write

$$V(x) = \left(\sqrt{V_s} \pm 2\left(\frac{2e^2 n_0}{\varepsilon\varepsilon_0}\right)^{1/2} x\right)^2 = V_s\left[1 \pm 2\left(\frac{2e^2 n_0}{\varepsilon\varepsilon_0 V_s}\right)^{1/2} x\right]^2. \tag{68.60}$$

Putting $\frac{1}{2}\left(\frac{\varepsilon\varepsilon_0 V_s}{2e^2 n_0}\right)^{1/2} = t$ we obtain

$$V(x) = V_s\left(1 \pm \frac{x}{t}\right)^2. \tag{68.61}$$

Since $V(x) > 0$ we retain the minus sign and obtain

$$V(x) = V_s\left(1 - \frac{x}{t}\right)^2. \tag{68.62}$$

It follows from this equation that *the electric field intensity dependence on the co-ordinate is linear*, and that the space charge density is constant. *A layer near the surface is depleted of the majority carriers and flooded with minority carriers.*

In conclusion consider the case of high fields $V > 0$ for which $\frac{V}{kT} \gg 1$

$$\frac{n_i^2}{N_d^2} e^{\frac{V}{kT}} \gg \frac{V}{kT} \gg 1. \tag{68.63}$$

According to (68.46) we have in this case

$$\frac{dV(x)}{dx} = \pm \left(\frac{2e^2kTn_0}{\varepsilon\varepsilon_0}\right)^{1/2} \frac{n_i}{N_d} e^{\frac{V}{2kT}} \tag{68.64}$$

and

$$V(x) = -2kT \ln\left[e^{-\frac{V_s}{kT}} + \frac{n_i}{n_0} \frac{x}{\sqrt{2}L^D}\right]. \tag{68.65}$$

Find the hole concentration for $x=0$:

$$p_s = p_0 e^{\frac{V_s}{kT}} = \frac{n_i^2}{N_d^2} N_d e^{\frac{V_s}{kT}} \gg N_d = n_0. \tag{68.66}$$

This means that in this case *the type of majority carriers in the surface layer changes sign, and the layer itself turns into an inversion layer*. Since the conductivity near the surface is opposite to that in the volume at some distance from the surface, a layer may be found where $p \cong n$, i.e. *an intrinsic conductivity*, or *i-type layer*. The region of the semiconductor where the conductivity changes from one type to the other is termed *physical p-n junction. It disappears after the external field has been switched off.*

The results of the analytical solution of the Poisson equation may be qualitatively interpreted as follows. The external electric field displaces the free charge carriers until it is compensated by their space charge. This means that when a negative potential is applied to a metal in contact with a semiconductor to which the positive potential is applied, the field is directed from the semiconductor to the metal and displaces the electrons from the surface into the volume of the semiconductor and the holes from the volume to the surface with the result that the surface layer becomes positively charged. A positive potential on the metal results in the surface layer being charged negatively. Depending on the conductivity type of the semiconductor this leads to flooding of the surface layer with majority carriers or to majority carrier depletion with subsequent inversion of the surface layer. If the field is applied instantaneously, the charge will be established in time equal to the Maxwell relaxation time. The arrangement shown in Fig. 101

and consisting of a metal electrode, a semiconductor and a dielectric (air) film between them may be regarded as a plane capacitor with the semiconductor as one of its electrodes. This enables the space charge evaluation to be made on the basis of the dielectric

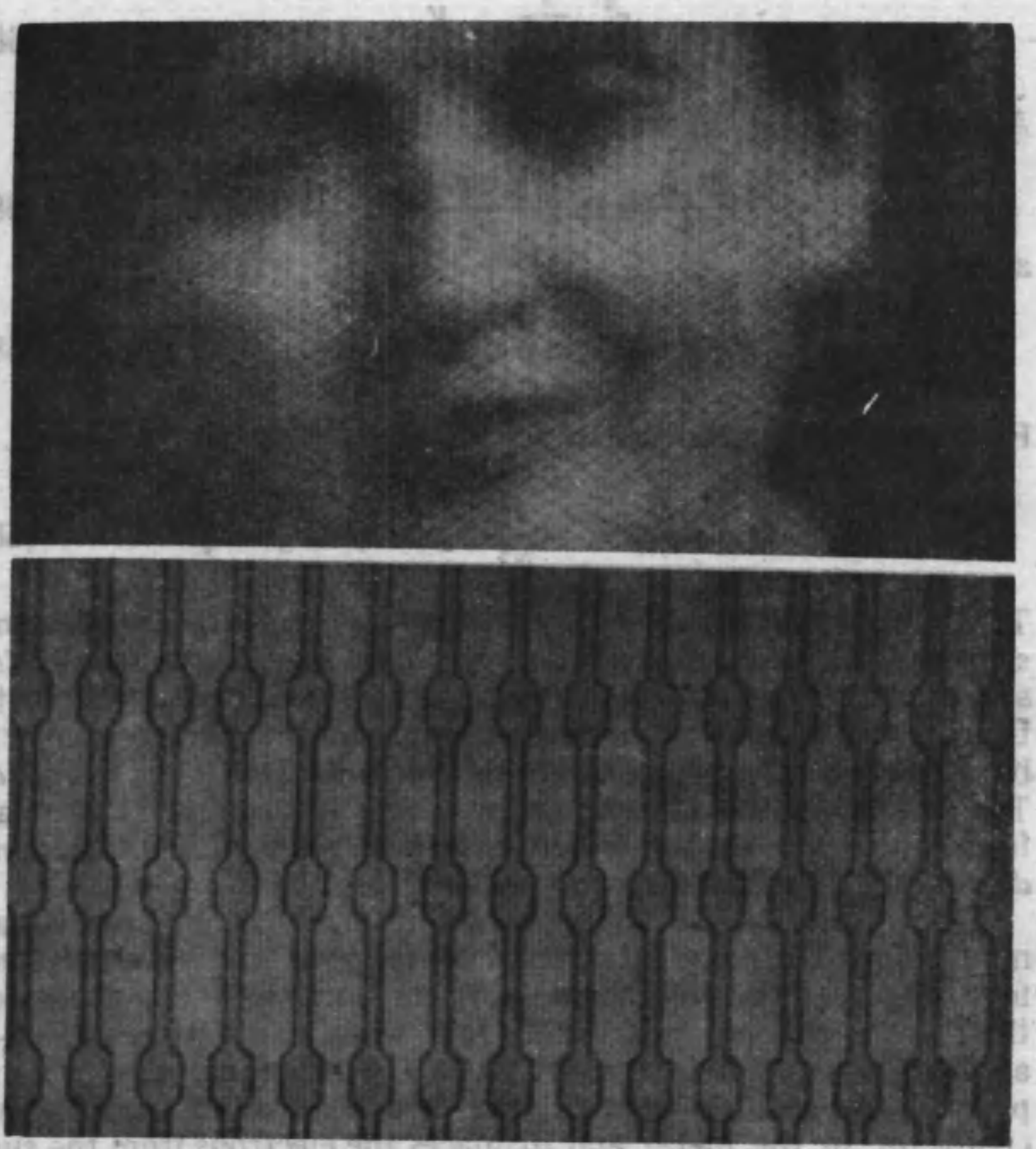


Fig. 103. The utilization of the field effect in silicon for image formation. Picture was imaged through an area device, of which part is shown here

film thickness and potential difference between the metal and the semiconductor. The space charge may be varied by varying the potential of the metal. The space charge pattern in the surface layer will reflect the pattern of the electric field set up by the electrode system *M*. By changing the external field pattern we are able to change the space charge pattern and the conductivity of the surface layer. The charge may be displaced along the surface either with the aid of additional electrodes or with the aid of a surface electrode system. This arrangement is the basis of one of the most important trends in solid state technology—that of MDS (metal-dielectric-semiconductor) devices which utilize thin dielectric films deposited on the semiconductor surface as dielectric layers. If the dielectric used is an oxide, the devices are termed MOS (metal-oxide-semiconductor) devices. Devices employing MDS structures are usually termed field-effect devices. The evolution of planar technology was responsible for widespread production of MDS devices.

The field-induced space charge may be established in conditions of thermodynamical equilibrium. It may, however, be established by non-equilibrium carriers as well. Indeed, if a space charge in the surface layer adjoining a contact is produced by illumination, it will persist at least for a time equal on the average to the lifetime, after which it will disappear owing to recombination. This means that the charge localized at the contact is able to store information about the light that created it at least during lifetime. By changing the pattern of the field one is able to make this localized charge move across the surface, and this makes scanning quite easy. Figure 103 shows an image produced by a vidicon utilizing the surface charge principle. Such devices have many advantages over the traditional vidicons utilizing photoconductivity.

Summary of Sec. 68

1. When a semiconductor is placed into a uniform external electric field the free charge carriers are displaced by this field, and a space charge is established which prevents the field from penetrating deep into the volume of the semiconductor. The electric field exists only inside the surface layer containing the space charge. The latter screens the external field and thereby prevents it from penetrating into the volume of the semiconductor.

2. The energy bands in the space-charge region are bent. If they are bent upwards, the surface layer will be flooded by holes, if downwards—by electrons. The conductivity of the surface layer in an intrinsic semiconductor is enhanced by the electric fields of both signs. In an extrinsic semiconductor the conductivity of the surface layer is enhanced if the external field extracts majority carriers, and is inhibited if the surface layer is flooded by minority carriers. The variations of surface layer conductivity occasioned by external fields are termed field effect.

3. The thickness of the layer which contains the space charge and the electric field and where the deflection of the energy bands

is noticeable is characterized by the quantity L^D:

$$L^D=\sqrt{\frac{\varepsilon\varepsilon_0 kT}{2e^2 n_i}};\quad L^D=\sqrt{\frac{\varepsilon\varepsilon_0 kT}{e^2 n_0}} \tag{68.1s}$$

for the intrinsic and the extrinsic semiconductor, respectively. L^D is termed the Debye length.

4. If the field intensity is sufficiently large, the conductivity type of the surface layer may change. The layer which changes its conductivity type is termed inversion layer. The layer in which $n\cong p\cong n_i$ is termed *i*-layer. The region where the conductivity changes its sign is termed physical *p-n* junction.

5. Should the surface states be occupied by free charge carriers, an electric field would be established between the surface and the volume of the semiconductor. This field would be localized in a surface layer the properties of which would be changed in the same way as by an external electric field.

6. Free electron concentration in metal is high. For this reason the Debye screening length is much less than the interatomic distance.

69. WORK FUNCTION

It is well known that work must be expended to transport an electron from a solid into vacuum. Assume the energy origin to be the energy of an electron immobile in relation to the body and placed in vacuum at an infinite distance from the body. In this case total energy of electrons at rest inside the body will be negative. The kinetic energy of the electron at the bottom of the conduction band is zero, and its total energy, from the standpoint of classical physics, is potential. Let the position of the bottom of the conduction band E_c in the "absolute" energy scale defined above be $-W$ ($W>0$). *W is equal to the work which must be performed to transport an electron resting inside the solid into vacuum without imparting kinetic energy to it. W is termed true work function.* In the theory of the quasifree electron ($-W$) was denoted by $\langle U\rangle$. The energy levels above E_c and up to the Fermi level are practically all occupied by electrons. Electrons with a negative total energy cannot leave the metal. But some of the electrons have a positive total energy; they can leave the metal. Calculate the electron flow from the metal into vacuum due to their thermal energy.

Let the metal fill the semi-space $x<0$. Calculate the number of electrons capable of negotiating a rectangular potential barrier of the height W and of leaving the metal. To this end the kinetic energy of the electron determined by its velocity v_x should be no less than the height of the potential well:

$$\frac{m^* v_x^2}{2} \geqslant W, \quad \frac{m^* v_x^2}{2} + E_c = E \geqslant 0. \tag{69.1}$$

It will be assumed that all the electrons that have left the metal are trapped by some external field and are unable to return into the metal. This assumption makes it easy to write an expression for the density of current flowing from the vacuum to the metal:

$$j_x = e \int_{v_{x\,min}}^{\infty} \iint_{-\infty}^{\infty} v_x 2 \frac{m^{*3}}{h^3} e^{-\frac{E-F}{kT}} dv_x dv_y dv_z =$$

$$= e \int_{v_{x\,min}}^{\infty} \iint_{-\infty}^{\infty} v_x e^{-\frac{E-F}{kT}} \frac{d\tau_p}{h^3}. \tag{69.2}$$

The expression (69.2) takes account of the fact that a unit volume of the crystal and the volume $d\tau_p$ contain $2\frac{1 \cdot d\tau_p}{h^3}$ states occupied by electrons with the probability

$$f_0(E, T) = \frac{1}{e^{\frac{E-F}{kT}} - 1} \cong e^{-\frac{E-F}{kT}}, \tag{69.3}$$

since for the electrons capable of leaving the metal $E - F \gg kT$.

The energy E may be expressed in terms of the velocity as follows:

$$E = E_c + \frac{m^* v^2}{2} = E_c + \frac{m^*}{2}(v_x^2 + v_y^2 + v_z^2), \tag{69.4}$$

and we obtain for j_x

$$j_x = \frac{2ee^{\frac{F-E_c}{kT}} m^{*3}}{h^3} \int_{-\infty}^{\infty} e^{-\frac{m^* v_y^2}{2kT}} dv_y \int_{-\infty}^{\infty} e^{-\frac{m^* v_z^2}{2kT}} dv_z \times$$

$$\times \int_{v_{x\,min}}^{\infty} v_x e^{-\frac{m^* v_x^2}{2kT}} dv_x. \tag{69.5}$$

The integrals over v_y and v_z may be reduced to the Poisson integral

$$\int_{-\infty}^{\infty} e^{-\xi^2} d\xi = \sqrt{\pi} \tag{69.6}$$

by substituting the variables, for example, as follows:

$$\frac{m^* v_y^2}{2kT} = \xi^2; \quad dv_y = \sqrt{\frac{2kT}{m^*}}\, d\xi. \tag{69.7}$$

Therefore

$$\int_{-\infty}^{\infty} e^{-\frac{m^* v_y^2}{2kT}} dv_y = \int_{-\infty}^{\infty} e^{-\frac{m^* v_z^2}{2kT}} dv_z = \sqrt{\frac{2\pi kT}{m^*}}. \tag{69.8}$$

Integration over v_x may be performed directly:

$$\int_{v_{x\,min}}^{\infty} e^{-\frac{m^* v_x^2}{2kT}} v_x\, dv_x = \frac{kT}{m^*} \int_{v_{x\,min}}^{\infty} e^{-\frac{m^* v_x^2}{2kT}} d\left(\frac{m^* v_x^2}{2kT}\right) =$$

$$= \frac{kT}{m^*} e^{-\frac{m^* v_{x\,min}^2}{2kT}} = \frac{kT}{m^*} e^{-\frac{W}{kT}} = \frac{kT}{m^*} e^{\frac{E_c}{kT}}. \tag{69.9}$$

Making use of (69.8) and (69.9) we write the expression for j_x:

$$j_x = \frac{4\pi e m^* k^2 T^2}{h^3} e^{\frac{F - E_c + E_c}{kT}} = AT^2 e^{\frac{F}{kT}}; \tag{69.10}$$

$$\left[A = \frac{4\pi e m^* k^2}{h^3} = 120 \left(\frac{m^*}{m}\right) \frac{A}{cm^2\, K^2} \right].$$

Denote the distance from the energy origin to the Fermi level F by Φ, i.e. put $F = -\Phi\ (\Phi > 0)$. Then j_x will be equal to

$$j_x = AT^2 e^{-\frac{\Phi}{kT}} = j_T. \tag{69.11}$$

Φ is termed thermodynamical electron work function of the metal. Numerically it is equal to the work which should be performed to withdraw an electron occupying the Fermi level from the metal. The expression (69.11) bears the name of Richardson's formula, and j_T is termed *thermionic emission current density*. Actually, special conditions should be created to enable the thermionic current density to reach the value of j_T. Indeed, the metal having lost some of its electrons becomes positively charged. The electric field established thereby retains the electrons and returns them to the metal. A dynamic equilibrium is established between the two opposite electron flows which turns j_x zero. In order to *maintain j_x at the constant level j_T, the electron cloud should be dispersed with the aid of an electric field. At the same time the loss of charge by the metal should be compensated.* Both these functions may be performed by the same field, as for instance this is being done in a vacuum diode. The magnitude of thermi-

onic emission depends sharply on the temperature. Indeed, for $\Phi = 2.5$ eV and $T \cong 300$ K the current density is $j_T \cong 10^{-36}$ A/cm^2, but if the temperature is increased five-fold ($T \cong 1500$ K) $j_T \cong 0.8$ A/cm^2, i.e. it increases 10^{36} times.

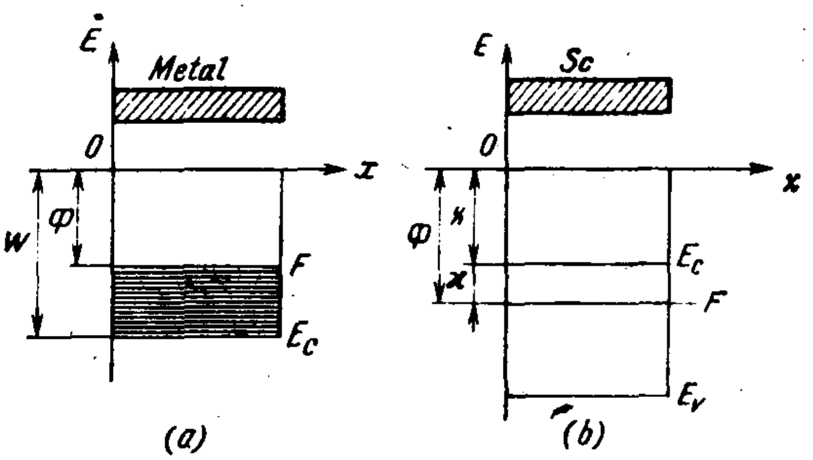


Fig. 104. A schematic diagram of energy levels in a metal (a) and a semiconductor (b). The energy of an electron at rest in vacuum is taken as the energy origin

Find the thermionic current density for electrons in a semiconductor. To this end consider the energy band diagram of the semiconductor. Let us assume as before the energy origin to be the energy of an electron resting in vacuum. Represent the Fermi energy in the form of a sum of two terms:

$$E_c = -\chi\,(\chi > 0) \qquad (69.11')$$

$$\varkappa = E_c - F, \qquad (69.12)$$

or in accordance with (69.12) and (69.11')

$$F = E_c - \varkappa = -(\chi + \varkappa). \qquad (69.13)$$

Such a representation is convenient because, as was demonstrated in Chapter III, χ depends only on the semiconductor type, while $\varkappa$ depends on doping and on temperature. Figure 104 shows the energy level diagram for a metal and for a semiconductor, the energy of an electron at rest in vacuum being taken as the origin of the energy scale.

Repeating the computations already performed for the metal, we obtain the density of the thermionic current from a semiconductor

$$j_T = AT^2 e^{\frac{F}{kT}} = AT^2 e^{-\frac{\chi+\varkappa}{kT}} = AT^2 e^{-\frac{\Phi}{kT}}, \qquad (69.14)$$

where Φ is *the thermodynamical work function of electrons in a semiconductor*. As in the case of a metal *it is determined by the position of the Fermi level in the "absolute" energy scale:*

$$\Phi = -F = \chi + \varkappa, \qquad (69.15)$$

despite the fact that there are no electrons on the Fermi level in a non-degenerate semiconductor. Write the expressions for the electron work function making use of the results of Chapter III.

For an intrinsic semiconductor the work function Φ is

$$\Phi(T)=\chi+\frac{\Delta E_0}{2}+\frac{kT}{2}\ln\frac{N_c}{N_v};\quad \Phi(0)=\chi+\frac{\Delta E_0}{2}. \qquad (69.16)$$

For an electron-type semiconductor

$$\Phi(T)=\chi+\frac{\Delta E_d}{2}+\frac{kT}{2}\ln\frac{N_c}{N_d}, \qquad (69.17)$$

and

$$\Phi(T)=\chi+kT\ln\frac{N_c}{N_d} \qquad (69.18)$$

in the high and low impurity ionization range, respectively.

For a hole-type semiconductor we have by analogy

$$\Phi(T)=\chi+\Delta E_0-\frac{\Delta E_a}{2}-\frac{kT}{2}\ln\frac{N_v}{N_a}, \qquad (69.19)$$

$$\Phi(T)=\chi+\Delta E_0-kT\ln\frac{N_v}{N_a}. \qquad (69.20)$$

It may be seen from (69.17-20) that the electron work function of the hole-type semiconductor is much larger than of the electron-type.

Summary of Sec. 69

1. The energy required to transport an electron of zero velocity from the metal into vacuum without changing its velocity is termed true work function W,

2. Thermodynamical electron work function of the solid is the term applied to the energy required to transfer an electron from the Fermi level to the level $E=0$ outside the crystal:

$$\Phi=-F. \qquad (69.1s)$$

3. The thermodynamical electron work function of a semiconductor depends on temperature and doping. The electron work function of a hole-type semiconductor exceeds that of an electron-type semiconductor by an amount almost equal to the forbidden band width.

4. The work function Φ determines the thermionic current density

$$j_T=AT^2e^{-\frac{\Phi}{kT}}. \qquad (69.2s)$$

70. CONTACT POTENTIAL DIFFERENCE. METAL-METAL CONTACT

Figure 105*a* shows the energy band diagram of two dissimilar metals M_1 and M_2 isolated from each other. The appropriate quantities relating to the two metals are indexed "1" and "2", respectively. The index "0" reflects the fact that the metals do not interact.

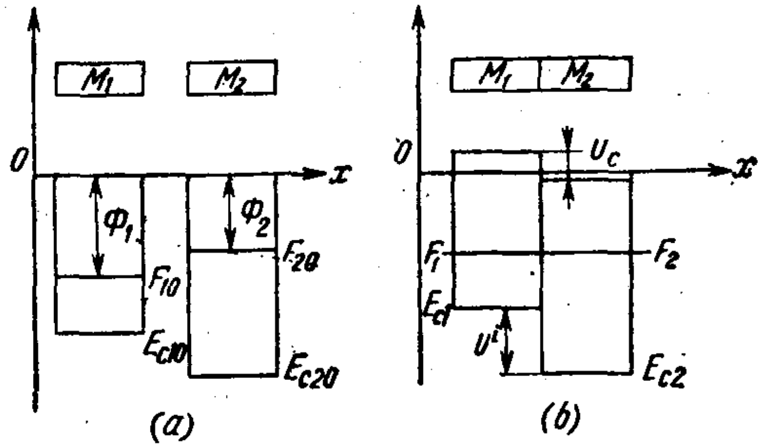


Fig. 105. A contact between two metals

Should the metals M_1 and M_2 be brought into contact they would start exchanging electrons. As is well known, particles go over from a subsystem with a higher Fermi level to a subsystem with a lower Fermi level. This may be demonstrated in the following way.

Consider two subsystems exchanging particles with each other. The variation of the system thermodynamic potential Φ is

$$d\Phi = \sum_i F_i dN_i = F_1 dN_1 + F_2 dN_2. \qquad (70.1)$$

However, since $dN_1 = -dN_2$, it follows

$$d\Phi = dN_1 (F_1 - F_2). \qquad (70.2)$$

If $F_1 = F_2$, $d\Phi = 0$ and the corresponding state is a state of equilibrium.

If, on the other hand, $F_1 \neq F_2$, then $d\Phi \neq 0$, and the system, as a consequence, is not in equilibrium. Transition to the equilibrium means that $d\Phi < 0$. For $F_1 < F_2$ the condition $d\Phi < 0$ requires $dN_1 > 0$, i.e. the number of particles of the first subsystem increases and that of the second decreases; for $F_1 > F_2$ $dN_1 < 0$. Thus, *the number of particles in the subsystem with higher chemical potential should decrease*. In other words, *the transition of a system to the equilibrium state is tantamount to the existence of a particle flow directed from the subsystem with a higher chemical potential to the subsystem with a lower chemical potential*. This directional particle flow continues until, owing to certain physical processes connected with the flow, the chemical potentials become equal.

The following simple reasoning may lead to the conclusion as to the nature of this directional particle flow. The electrons should occupy all the lower free levels; the Fermi level separates the mainly occupied levels from the almost free; therefore, if F_{20} lies above F_{10}, the electrons from higher M_2 levels should go over to lower free M_1 levels in full accord with the result following from the condition $d\Phi \leqslant 0$.

Since the Fermi level position in the "absolute" (i.e. valid for all bodies) energy reference system is determined solely by the thermodynamical work function Φ, we may formulate the following conclusion: *when two bodies are brought into contact, a directional particle flow from the body with a smaller work function to the body with the greater work function is initiated.* The directional flow will cease when the energy level population of both subsystems will be equalized. The necessary number of electrons may, however, turn out to be quite small. Indeed, as electrons leave the body it becomes positively charged, and its electrons acquire an additional negative energy with the result that all the energy levels are lowered by a certain amount. The energy levels of the second metal which acquires a negative charge rise by some other amount, generally, not equal to the first. The directional electron flow from the body with lower work function to the body with greater work function will cease the moment the Fermi levels of both bodies become equal: $F_1 = F_2$. It must, however, be kept in mind that $F_1 = F_2$ are not identical with both F_{10} and F_{20}. Figure 105*b* shows the energy level pattern of two contacting metals in equilibrium.

Since the bodies become charged, a potential difference and an electric field is established between them. As was demonstrated in Sec. 68, in a metal the electric field is localized in a thin layer the thickness of which is much less than the lattice constant.

The charge acquired by the metal will, generally, be distributed over its surface. Therefore, the electric field will in this case be localized in a thin surface layer of the metal. A potential difference φ^c termed *external contact potential difference* should exist between any two points M_1 and M_2 on the outer surface of the metals. The value of this contact potential difference may be obtained with the help of simple reasoning: the energy levels in M_1 and M_2 rise and drop, so that their displacement (as equilibrium is attained) becomes equal to the difference of the respective work functions, therefore

$$e\varphi_{21}^c = U_{12}^c = \Phi_1 - \Phi_2. \tag{70.3}$$

U^c represents the potential barrier which provides for the equality of the thermodynamical work functions for the transport of electrons from M_1 to M_2 and from M_2 to M_1 required by the energy conservation law. Indeed, take an electron belonging to the metal and having a total energy E and make it move round a closed circuit in opposite directions (Fig. 106). The total energy of the electron in the metals M_1 and M_2 remains constant and is equal to E. When moving counterclockwise take the electron out of M_2 into the vacuum. Its energy will be $E+\Phi_2$. Insert it into M_1; this will decrease its energy by the amount Φ_1: $E^{(1)}=E+\Phi_2-\Phi_1$. Now make the trip in the clockwise direction. When the electron returns to M_1 it will have the energy $E^{(2)}=E+\Phi_1-\Phi_2$. The energy variation in the first case is $\Delta E^{(1)}=E^{(1)}-E=\Phi_2-\Phi_1$, and in the second case, $\Delta E^{(2)}=E^{(2)}-E=\Phi_1-\Phi_2$. Since $\Phi_1\neq\Phi_2$ the circulation in one direction increases the energy and in the other direction decreases the energy in contradiction with the energy-conservation law. This was because we ignored the energy U^c produced by the external contact potential difference φ^c which should be added to the total energy, or subtracted from it depending on the direction of circulation. With U^c accounted for, we obtain

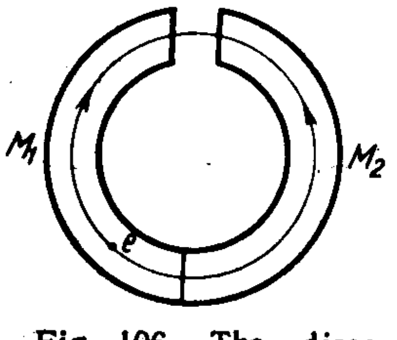


Fig. 106. The direction of closed circuit circulation

$$\left.\begin{aligned}\Delta E^{(1)}&=\Phi_2+U^c-\Phi_1=0,\\ \Delta E^{(2)}&=\Phi_1-(\Phi_2+U^c)=0.\end{aligned}\right\}\qquad(70.4)$$

It follows from (70.4) that the magnitude of the potential barrier U^c dividing two arbitrary outside surface points of the two metals M_1 and M_2 is equal to the difference of the thermodynamical work functions. Physically this means that the potential of the metal with the lower work function is raised in relation to the metal with the higher work function by the amount φ^c, with the result that the work function at any point of the metallic pair will be determined by the larger of the two work functions Φ_1 or Φ_2, and the external contact field U^c will be localized in the immediate vicinity of the metal with the lower work function.

Table 22 shows by way of example the values of the electron work function Φ for some solids.

The equality of thermionic emission current densities is provided for by external contact potential difference which is due to the difference in electron work functions of the metals.

External contact potential difference between metals may be established not only when they are in contact but also when *the metals can exchange electrons via termionic emission*, there being no direct contact between them.

Table 22

Material	Li	Be	C	Na	K	Sn	W	Pt	Cs
F, eV	4.71	9.07	11.05	3.07	2.04	4.07	5.81	6.00	1.53
Φ, eV	2.39	3.92	4.62	2.35	2.22	4.3	4.52	5.36	1.93

Now we will consider the conditions prevailing in the contact itself. If in equilibrium conditions the Fermi levels of the metals M_1 and M_2 coincide it means that there is a potential barrier at the contact and an electric field established by it. This field E^c *is termed contact field*, and the corresponding *potential difference* φ^i *is termed internal contact potential difference.* It may be found from the condition

$$\varphi^i_{21}=\frac{U^i_{12}}{e}, \tag{70.5}$$

where U^i_{12} is the potential barrier between M_1 and M_2. It may be seen directly from Fig. 105*b* that to transfer an electron with zero kinetic energy from the first metal to the second without changing its velocity the energy should be expended equal to the difference in the energy of the electron in positions E_{c1} and E_{c2}. Therefore the value of the potential barrier is

$$U^i_{12}=E_{c1}-E_{c2}. \tag{70.6}$$

U^i_{12}, on the other hand, may be expressed in terms of kinetic energy variations of the particle crossing the contact:

$$U^i_{12}=T_2-T_1. \tag{70.7}$$

The conditions (70.7) and (70.6) coincide because $T_1=E-E_{c1}$; $T_2=E-E_{c2}$. But since all the levels were displaced by the amount U^c it follows that

$$E_{c1}-E_{c2}=E_{c10}-E_{c20}+U^c_{12}=E_{c10}-E_{c20}+\Phi_1-\Phi_2. \tag{70.8}$$

Taking into account that $\Phi_1=-F_{10}$, $\Phi_2=-F_{20}$ we obtain

$$\begin{aligned}U^i_{12}&=E_{c1}-E_{c2}=(E_{c10}-F_{10})-(E_{c20}-F_{20})=\\&=(F_{20}-E_{c20})-(F_{10}-E_{c10}).\end{aligned} \tag{70.9}$$

Hence, *the potential barrier U^i_{12} determining the internal contact potential difference $\varphi^i=\frac{U^i}{e}$ is equal to the difference between the Fermi levels of isolated metals M_1 and M_2 measured from the bottom of the conduction bands of the respective metals.* The value of the internal potential barrier is determined by the electron concentration in the metals. Since

$$F_0-E_{c0}=\frac{h^2}{2m^*}\left(\frac{3n}{8\pi}\right)^{2/3}, \tag{70.10}$$

it follows

$$U^i_{12}=\frac{h^2}{2}\left(\frac{3}{8\pi}\right)^{2/3}\left(\frac{n_2^{2/3}}{m_1^*}-\frac{n_1^{2/3}}{m_2^*}\right). \tag{70.11}$$

The contact field U^i provides for the equality of the electron currents from one metal to the other and vice versa. Indeed, write the expressions for the density of the electron flows from the first metal into the second $\left(\frac{j_{21}}{e}\right)$ and from the second into the first $\left(\frac{j_{12}}{e}\right)$:

$$\frac{j_{21}}{e}=\frac{2m_1^{*3}}{h^3}\int\limits_{v_{x1}}^{\infty}\iint\limits_{-\infty}^{\infty}\frac{v_x\,dv_x\,dv_y\,dv_z}{e^{\frac{E_1-F_1}{kT}}+1}, \tag{70.12}$$

$$\frac{j_{12}}{e}=\frac{2m_2^{*3}}{h^3}\int\limits_{v_{x2}}^{\infty}\iint\limits_{-\infty}^{\infty}\frac{v_x\,dv_x\,dv_y\,dv_z}{e^{\frac{E_2-F_2}{kT}}+1}. \tag{70.13}$$

The expression (70.9) may be obtained from the condition of the equality of the flows (70.12) and (70.13) and vice versa: the equality of the flows $\frac{j_{21}}{e}=\frac{j_{12}}{e}$ follows from (70.9) or (70.7)

Table 22 shows the values of the Fermi energy of some metals which may be used to calculate U^i.

71. METAL-SEMICONDUCTOR CONTACT

Consider now a metal-semiconductor contact. Figure 107*a* shows the energy band diagram of a metal and a semiconductor the instant they have been brought into contact. This is a non-equilibrium state, and in the process of electron exchange the electron flow from the substance with higher Fermi energy into the substance with lower Fermi energy will prevail. If the Fermi level of the metal F_m lies below the Fermi level of the semiconductor F_{sc} $(F_m < F_{sc})$, the electrons will go over from the semiconductor to the metal. The metal will acquire a negative, and the semiconductor, a positive charge. The directional electron flow will continue until the Fermi levels are equal, and a dynamic equilibrium is thus established. If $F_m > F_{sc}$, the electrons will leave the metal for the semiconductor. The electron exchange can take place not only in the case of direct contact between the metal and the semiconductor but via thermionic emission, which leads to the establishment of an external contact potential difference $e\varphi^c_{msc} = U^c_{scm} = \Phi_{sc} - \Phi_m$ as well. When two metals are in contact φ^c is distributed approximately equally over both metals since their electron concentrations are of about the same magnitude. In each metal the potential is constant. This is not the case with a semiconductor. The potential at different points in the semiconductor may assume different values

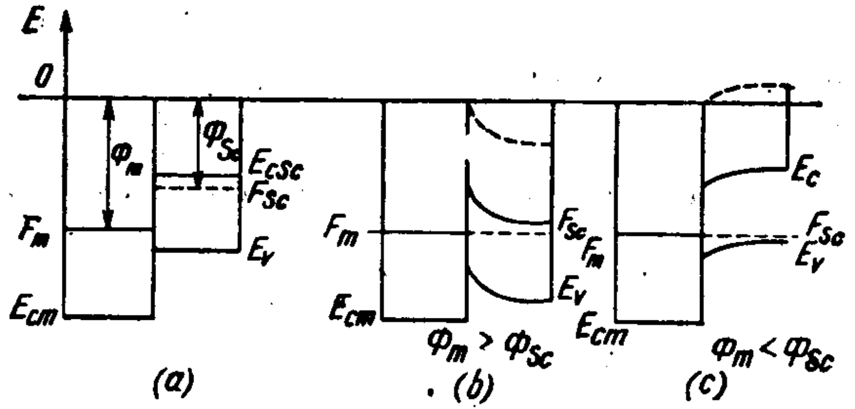


Fig. 107. Metal-semiconductor contact

since the external field may penetrate deep into the semiconductor producing space charges and bending the energy bands. Practically *the entire potential difference will be concentrated inside the contact layer of the semiconductor deflecting the energy bands in this layer* (Fig. 107*b*, *c*). For $\Phi_{sc} < \Phi_m$ the deflection will be upwards, and for $\Phi_{sc} > \Phi_m$ it will be downwards. But the deflection of the energy bands in the contact layer changes the electron and hole concentrations: an upward deflection of the energy bands increases the hole concentration, and a downward deflection does the same to the electron concentration. Thus, *if the electron work function of the semiconductor is less than that of the metal, the contact layer will be flooded by holes. If, on the other hand, $\Phi_{sc} > \Phi_m$, it will be flooded by electrons.*

If the energy bands are bent in an intrinsic semiconductor the conductivity of the contact layer will be increased. This is not the case with extrinsic conductivity. The conductivity is increased when the contact layer is flooded by the majority carriers and it is decreased when it is flooded by the minority carriers. *The enhanced conductivity layer (flooded by majority carriers) is termed antibarrier layer. The depleted conductivity layer (flooded by minority carriers) is termed barrier layer. The layer is termed inversion layer if a large energy band deflection results in the layer changing its conductivity type.*

Table 23 shows the properties of the barrier, antibarrier, and inversion layers.

Table 23

Semiconductor type	Antibarrier layer	Barrier layer	Conductivity type inversion
Electron	$F_m > F_{sc}$; $\Phi_m < \Phi_{sc}$	$F_m < F_{sc}$; $\Phi_m > \Phi_{sc}$	$F_m < F_{sc}$; $\Phi_m \geqslant \Phi_{sc} + \Delta E_0$
Hole	$F_m < F_{sc}$; $\Phi_m > \Phi_{sc}$	$F_m > F_{sc}$; $\Phi_m < \Phi_{sc}$	$F_m > F_{sc}$; $\Phi_m \leqslant \Phi_{sc} - \Delta E_0$

The flooding of the semiconductor contact layer by minority carriers plays an important part in all physical measurements in semiconductors utilizing electric currents. Indeed, if an external voltage is applied to the circuit consisting of a metal and a semiconductor, a current will be established in it consisting of an electron and a hole current:

$$\mathbf{j} = \mathbf{j}_n + \mathbf{j}_p. \qquad (71.1)$$

Since the electron and hole concentrations at different points of the semiconductor vary, *the composition of the current varies* too. Its total value remains constant. To be definite suppose that the semiconductor is of the electron-type. Denote the fraction of the current carried by *the minority carriers* (holes in this case) by $\gamma(x)$:

$$\gamma(x) = \frac{j_p}{j} = \frac{j_p(x)}{j_n(x) + j_p(x)} = \frac{p(x)}{p(x) + bn(x)}; \quad \left(b = \frac{|\mu_n|}{\mu_p}\right). \qquad (71.2)$$

Denote the value of γ corresponding to equilibrium minority carrier concentration by γ_0:

$$\gamma_0 = \frac{p_0}{p_0 + bn_0}. \qquad (71.3)$$

The current flow is accompanied by a decrease in minority carrier concentration at the point x for $\frac{\gamma(x)}{\gamma_0(x)} < 1$ and an increase for $\frac{\gamma(x)}{\gamma_0(x)} > 1$.

In other words, *the current flowing through the contact region changes the concentrations of the majority and the minority carriers in the volume of the semiconductor with the resultant change in its properties*. If $\frac{\gamma(x)}{\gamma_0(x)} > 1$, the semiconductor volume will be flooded by minority (non-equilibrium) carriers. This will be possible both for the barrier and the antibarrier layers. In the first case the resultant process is carrier injection, in the second—accumulation. If $\frac{\gamma(x)}{\gamma_0(x)} < 1$, the volume will be depleted of the minority carriers. If this depletion is linked with the existence of a barrier layer, the phenomenon is termed exclusion; in case of an antibarrier layer the term applied is extraction. These phenomena correspond to those described in Sec. 66. In this section we have revealed the cause of non-equilibrium minority carrier concentrations—the bending of the energy bands by the contact field. The external electric field makes the charge carriers drift, and this leads to the phenomena of carrier injection, extraction, accumulation, and exclusion.

A contact which does not introduce any changes in the minority carrier concentration ($\gamma=\gamma_0$) *is termed ohmic.* The provision of ohmic contacts for the investigation of physical phenomena in semiconductors is a very important and at the same time a very difficult problem. Since it is not always possible to make an ohmic contact, *semiconductor physics makes a wide use of compensation measurement methods to exclude the effects of contact phenomena on the results of the measurements.* In these methods no current passes through the contacts, and for this reason there are no contact phenomena to distort the results of the measurements.

Table 24 lists the conditions for studying contact phenomena for various types of contacts, semiconductor conductivity type and field direction in the assumption that the contact ($x=0$) is on the left boundary of the semiconductor.

The investigations into the properties of point contacts to germanium and silicon carried out as far back as the fifties have, however, demonstrated that in some cases the nature of the metal does not affect these properties. In other words, it would seem that the picture of band deflection leading to the flooding of the contact layer by (or to its depletion of) the minority carriers does not correspond to the facts. The explanation was found when the effects of surface states on the volume properties of semiconductors were considered, as was discussed in Sec. 68. We learned that the existence of the surface states occasioned the bending of the energy bands in the surface layer. Should this bending be large (should V_s be large) the additional bending occasioned by the metal-semiconductor contact would little affect the properties of such contacts. In other words, *if the barrier or antibarrier layers are actually produced by the surface states, their properties will little depend on the nature of the contact metal.* The existence of the surface states makes itself manifest in the decrease of the carrier lifetimes as well, as was discussed in Sec. 67. The surface states are classified into the "slow" and the "fast" depending on the time it takes for the carrier exchange cycle between them and the

Table 24

Contact	Rectifying (barrier layer)	Non-rectifying (antibarrier layer)	Ohmic
	$\frac{\gamma(0)}{\gamma_0} \neq 1$	$\frac{\gamma(0)}{\gamma_0} \neq 1$	$\frac{\gamma(0)}{\gamma_0} = 1$
Injection	$\gamma(x) > \gamma_0(x)$ $E > 0$, (n-type) $E < 0$, (p-type)	—	—
Exclusion	$\gamma(x) < \gamma_0(x)$ $E < 0$, (n-type) $E > 0$, (p-type)	—	—
Extraction	—	$\gamma(x) < \gamma_0(x)$ $E > 0$, (n-type) $E < 0$, (p-type)	—
Accumulation	—	$\gamma(x) > \gamma_0(x)$ $E < 0$, (n-type) $E > 0$, (p-type)	—

semiconductor volume to be completed. The relaxation time for the "*slow*" *states* varies from tenths of a second to many hours. The corresponding values for the "*fast*" *states* are from 10^{-6} to 10^{-4} s. The "slow" states are created by mobile atoms or ions adsorbed on the oxide film or on some other film on the surface of the semiconductor. The relaxation time of the "slow" states involves not only the carrier trapping time but also the time it takes for the atom or ion to move in the field and thereby create or destroy the state. The "fast" states are purely electron states localized on the boundary separating the semiconductor volume from the oxide film.

Summary of Secs. 70-71

1. A potential difference φ^c termed external contact potential difference exists between any two points on the surfaces of two metals in contact with each other. The discontinuity in potential energy $U^c = e\varphi^c$ provides for the equality of the thermionic electron currents from one metal to the other and vice versa. It follows from the energy conservation law or from the condition of equivalence of thermionic electron currents that

$$U_{12}^c = e\varphi_{12}^c = \Phi_1 - \Phi_2. \qquad (70.1s)$$

It follows from the condition of equivalence of electron currents passing in opposite directions through a contact that there is a potential difference φ^i across the contact. This is termed internal contact potential difference, and the discontinuity in potential energy U^i is

$$U^i_{12} = e\varphi^i_{21} = T_2 - T_1 = (F_{2} - E_{c20}) - (F_{10} - E_{c10}). \qquad (70.2s)$$

2. Since an electric field cannot exist in the volume of a metal in the absence of current, the contact potential difference is concentrated entirely in the contact layer of the semiconductor. This results in the deflection of the energy bands and in the appearance of a space charge in the same way as when an external electric field is applied to the semiconductor. The energy band deflection resulting from the exchange of electrons via thermionic emission will not be altered when the metal and the semiconductor are brought into contact. Since there is a potential barrier of the height

$$U^c_{msc} = e\varphi^c_{scm} = \Phi_m - \Phi_{sc}, \qquad (70.3s)$$

preventing the electrons from going over from the semiconductor to the metal, it may be said that at the metal-semiconductor contact an internal contact potential difference is established which is equal to the external contact potential difference, i.e. to the difference between the work functions of the metal and the semiconductor.

3. If the contact layer of the semiconductor is flooded by majority carriers, it is termed antibarrier layer. If it is flooded by minority carriers, the corresponding term is barrier layer. If the flooding by minority carriers is so great that the conductivity type of the layer is changed, it is termed inversion layer and there is a physical *p-n* junction lying underneath it. A large deflection of the bands accompanied by flooding of the contact layer by majority carriers may bring about the degeneracy of the semiconductor inside the layer.

4. As current passes through a metal-semiconductor contact, minority carriers may be drawn by the field into regions where they may become non-equilibrium carriers. This leads to the phenomena of injection, extraction, exclusion and accumulation, affecting the properties of the semiconductor. This is the reason why compensation measurement methods are used in semiconductor physics.

5. A contact which injects or extracts minority carriers is termed rectifying.

6. A contact which excludes or accumulates minority carriers is termed non-rectifying.

7. A contact is termed ohmic if the composition of the current is the same in the contact and in the volume of the semiconductor.

8. When two bodies are brought into contact the electrons start going over from the body with the higher Fermi energy into the body with the lower Fermi energy. Directional flow ceases as soon as the Fermi levels become equal.

72. INHOMOGENEOUS SEMICONDUCTOR. p-n JUNCTION

There are numerous semiconductor devices which utilize the properties of the p-n junction. To gain insight into these properties we may consider the contact between an electron and a hole-type sample of the samè semiconductor. However, another approach based on the study of inhomogeneous semiconductors is also possible. Suppose we have a semiconductor sample with donor N_d (**r**) and acceptor N_a (**r**) impurities arbitrarily distributed in it. This results in electron and hole concentrations being dependent on the co-ordinate. The position of the Fermi level in respect to the energy bands E_v and E_c is determined by the electron or hole concentrations, therefore $F=F(\mathbf{r})$. But if $F=F(\mathbf{r})$, the corresponding state is a non-equilibrium state entailing a charge carrier flow directed so as to equalize the electron and hole concentrations. *The diffusion current* $\mathbf{j}_D$ *results in charge separation leading to the creation of a space charge and of the accompanying electric field which bends the energy bands.* In thermodynamical equilibrium the Fermi level is independent of the co-ordinate: $F = F_0$. The diffusion current $\mathbf{j}_D$ is compensated by the drift current $\mathbf{j}_E$, therefore

$$\mathbf{j}=\mathbf{j}_D+\mathbf{j}_E=0. \tag{72.1}$$

We may use (72.1) to assess the intensity of the built-in electric field $\mathbf{E}^i$:

$$\mathbf{E}^i=-\frac{\mathbf{j}_D}{\sigma}=-\frac{\mathbf{j}_{D_n}+\mathbf{j}_{D_p}}{\sigma_n+\sigma_p}, \tag{72.2}$$

$$\mathbf{E}^i(\mathbf{r})=\frac{D_p\nabla p-D_n\nabla n}{\mu_p p-\mu_n n}=\frac{D_p(\nabla p-b\nabla n)}{\mu_p(p+bn)}. \tag{72.3}$$

The field intensity $\mathbf{E}^i$, as may be seen, is determined by the electron and hole concentration gradients and by the conductivity. Maximum values of the field $\mathbf{E}^i$ are attained when the conductivity is at its minimun and when the directions of the electron and hole concentration gradients are opposite. The field $\mathbf{E}^i$ (**r**) may be expressed in terms of the concentration gradient of the charge carriers of one type. We shall confine ourselves to the case of non-degenerate semiconductor for which $np=n_i^2$ and

$$p\nabla n+n\nabla p=0. \tag{72.4}$$

Substituting ∇n from (72.4) into (72.3) we obtain

$$\mathbf{E}^i(\mathbf{r})=\frac{D_p}{\mu_p}\frac{p+bn}{p+bn}\frac{\nabla p}{p}=\frac{kT}{e_p}\frac{\nabla p}{p}=\frac{kT}{e_n}\frac{\nabla n}{n}. \tag{72.5}$$

The relation (72.5) may be written in the following form:

$$\frac{\nabla n}{n}=\nabla\ln n=-\frac{1}{kT}\nabla(E_c-F) \tag{72.6}$$

and

$$\mathbf{E}^i(\mathbf{r}) = -\frac{1}{e_n}\nabla(E_c - F) = -\frac{\nabla E_c}{e_n} = \frac{\nabla E_v}{e_p}. \tag{72.7}$$

Consider a semiconductor doped with impurity of one type, for instance, with donors $N_d = N_d(\mathbf{r})$.

In case the impurity is completely ionized $N_d^+ = N_d$, and the temperature is in the extrinsic range,

$$\mathbf{E}^i = \frac{kT}{e_n}\frac{\nabla N_d(\mathbf{r})}{N_d(\mathbf{r})} = \frac{kT}{e_n}\nabla \ln N_d(\mathbf{r}). \tag{72.8}$$

Suppose the impurity concentration $N_d(\mathbf{r})$ varies exponentially:

$$N_d(\mathbf{r}) = N_d(\mathbf{r}_0)\, e^{-(\varkappa,\, \mathbf{r} - \mathbf{r}_0)}. \tag{72.9}$$

Then the electric field intensity $\mathbf{E}^i$ will be determined by the quantity $\varkappa$:

$$\mathbf{E}^i(\mathbf{r}) = -\frac{kT}{e_n}\varkappa. \tag{72.10}$$

Here $\varkappa^{-1}$ is numerically equal to the distance at which the impurity concentration changes e times. For $\varkappa = 1\ \text{cm}^{-1}$ $\mathbf{E}^i = 0.026$ V/cm; for $\varkappa = 10^7\ \text{cm}^{-1}$ $\mathbf{E}^i = 0.026 \times 10^7 = 2.6 \times 10^5$ V/cm ($T \cong 300$ K).

These results are also valid when the semiconductor is doped with impurities of both types provided one of them is distributed uniformly, and the other non-uniformly. For example, should a hole-type semiconductor with a hole concentration $p_0 = N_a^+ = N_a$ be doped with a donor impurity with an arbitrary concentration distribution, the field intensity $\mathbf{E}^i$ at points where the conductivity retains its original hole type could be found with the aid of expression (72.8). The same expression will also hold in the region where $N_d > N_a$. In the compensation region the Poisson equation of the type (68.12) should be used.

Consider one specific case when the impurity is distributed uniformly for $x < 0$ and for $x > 0$ but changes its type at $x = 0$:

$$N_d(x) = \begin{cases} N_d & \text{for } x < 0 \\ 0 & \text{for } x > 0, \end{cases}$$
$$N_a(x) = \begin{cases} 0 & \text{for } x < 0 \\ N_a & \text{for } x > 0. \end{cases} \tag{72.11}$$

To the left of the boundary $\delta p_a = 0$, $n_0 = N_d$; to the right, $\delta n_d = 0$, $p_0 = N_a$. Moreover, away from the origin of co-ordinates $\delta n = \delta p = 0$, i.e. $V(x) = 0$. In other words, *there is no space charge and no electric field away from the plane* $x = 0$. Consider the vi cinity of the plane $x=0$. Write the equation (68.12) for $x<0$

$$\frac{d^2V}{dx^2}=-\frac{e^2}{\varepsilon\varepsilon_0}\left\{N_d\left(e^{-\frac{V(x)}{kT}}-1\right)-\frac{n_i^2}{N_d}\left(e^{\frac{V(x)}{kT}}-1\right)+\delta n_d\right\}, \quad (72.12)$$

and for $x>0$,

$$\frac{d^2V}{dx^2}=\frac{e^2}{\varepsilon\varepsilon_0}\left\{N_a\left(e^{\frac{V(x)}{kT}}-1\right)-\frac{n_i^2}{N_a}\left(e^{-\frac{V(x)}{kT}}-1\right)+\delta p_a\right\}. \quad (72.13)$$

To solve the equations (72.12) and (72.13) δp_a and δn_d should be expressed in terms of $V(x)$. It may, however, be easily demonstrated that these equations may be simplified. First determine the sign of $V(x)$. Some electrons from the n-region will go over to the p-region, and some of the holes from the p-region will go over to the n-region, charging the n-region positively, and the p-region negatively.

An electric field will be established in the p-n junction area directed from the n-region to the p-region. Since the electron potential energy gradient and the electric field coincide in direction, we may say that *in the vicinity of the plane $x=0$ to the left of it the energy bands are deflected upwards in respect to their position in the volume*, i.e. that $V(x)>0$ for $x<0$. In the p-region the energy bands are deflected downwards in respect to their position away from $x=0$, i.e. $V(x)<0$ for $x>0$. The rise of the energy levels for $x<0$ may bring about a change in the number of electrons occupying the donor level. We, however, will consider the case when such a change may be neglected, i.e. we assume $\delta n_d=0=\delta p_a$.

We may write for the area where $|V(x)|>kT$,

$$\frac{d^2V(x)}{dx^2}=\frac{e^2}{\varepsilon\varepsilon_0}\left(N_d+\frac{n_i^2}{N_d}e^{\frac{V(x)}{kT}}\right) \qquad (x<0), \quad (72.14)$$

$$\frac{d^2V(x)}{dx^2}=-\frac{e^2}{\varepsilon\varepsilon_0}\left(N_a+\frac{n_i^2}{N_a}e^{-\frac{V(x)}{kT}}\right)(x>0). \quad (72.15)$$

If the doping level of the semiconductor is sufficiently high, so that

$$\frac{n_i^2}{N_d^2}e^{\frac{V(x)}{kT}}\ll 1 \text{ and } \frac{n_i^2}{N_a^2}e^{-\frac{V(x)}{kT}}\ll 1, \quad (72.16)$$

the result will be

$$\begin{aligned}&\rho^+(x)=e^+N_d^+=e^+N_d && \text{for } x<0, \text{ and}\\ &\rho^-(x)=e^-N_a^-=-e^+N_a && \text{for } x>0,\end{aligned}$$

i.e. *a space charge of constant density* equal to the ion charge density will be established in the region $x=0$. Suppose that the region of the space charge is bounded by the points $(-t_n, 0)$ and $(0, t_p)$. The field intensity may be found from (72.14-16):

$$\frac{dV(x)}{dx}=\frac{e^2}{\varepsilon\varepsilon_0}N_d x+C_1 \quad (x<0), \tag{72.17}$$

$$\frac{dV(x)}{dx}=-\frac{e^2}{\varepsilon\varepsilon_0}N_a x+C_2 \quad (x>0). \tag{72.18}$$

C_1 and C_2 may be found from the condition that the field turns zero at the points $x_1=-t_n$ and $x_2=t_p$:

$$C_1=\frac{e^2}{\varepsilon\varepsilon_0}N_d t_n; \quad \frac{dV(x)}{dx}=\frac{e^2 N_d}{\varepsilon\varepsilon_0}(t_n+x), \quad (-t_n<x<0) \tag{72.19}$$

$$C_2=\frac{e^2}{\varepsilon\varepsilon_0}N_a t_p; \quad \frac{dV(x)}{dx}=\frac{e^2 N_a}{\varepsilon\varepsilon_0}(t_p-x), \quad (0<x<t_p). \tag{72.20}$$

The condition of the field continuity at the point $x=0$ yields

$$N_d t_n=N_a t_p, \tag{72.21}$$

i.e. *the thickness of the space charge layer in each region is inversely proportional to its impurity concentration*. Denoting the combined space charge layer thickness by $t=t_n+t_p$ we may write

$$t_p=\frac{N_d}{N_a+N_d}t; \quad t_n=\frac{N_a}{N_a+N_d}t. \tag{72.22}$$

Hence, with regard to (72.19) and (72.20) we may say that *the electric field intensity in the space charge region is a linear function of the co-ordinate*. Integrating (78.19-20) over x we obtain the co-ordinate dependence of the potential energy in the form of a *quadratic parabola*:

$$V(x)=\frac{e^2 N_d}{2\varepsilon\varepsilon_0}(t_n+x)^2+C_1' \tag{72.23}$$

and

$$V(x)=-\frac{e^2 N_a}{2\varepsilon\varepsilon_0}(t_p-x)^2+C_2'. \tag{72.24}$$

Apply the following reasoning to find C_1' and C_2'. Put $V(x)=0$ for $x<-t_n$ and $V(x)=U^c$ for $x>t_p$, i.e. the deflection of the energy bands will be measured from their position in the volume of the n-region. *U^c is the height of the potential barrier for electrons established at the boundary of the n- and p-regions*:

$$V(x)=\frac{e^2 N_d}{2\varepsilon\varepsilon_0}(t_n+x)^2=V_n(x) \quad (x<0), \tag{72.25}$$

$$V(x)=-\frac{e^2 N_a}{2\varepsilon\varepsilon_0}(t_p-x)^2+U^c=V_p(x) \quad (x>0). \tag{72.26}$$

Putting $x=0$ and taking into account that $V(x)$ is continuous we obtain

$$V_n(0)=V_p(0);\quad \frac{e^2N_d}{2\varepsilon\varepsilon_0}t_n^2=-\frac{e^2N_a}{2\varepsilon\varepsilon_0}t_p^2+U^c, \tag{72.27}$$

or making use of (72.22),

$$U^c=\frac{e^2}{2\varepsilon\varepsilon_0}(N_dt_n^2+N_at_p^2)=\frac{e^2N_dt_n}{2\varepsilon\varepsilon_0}(t_n+t_p)=\frac{e^2N_d^2t_n^2}{2\varepsilon\varepsilon_0}\left(\frac{1}{N_d}+\frac{1}{N_a}\right). \tag{72.28}$$

The height of the potential barrier U^c may be expressed in terms of the difference between the distances $F-E_c$ in the n- and p-regions,

$$U^c=E_c(x>t_p)-E_c(x<-t_n)=\varkappa_p-\varkappa_n=$$
$$=(\chi+\varkappa_p)-(\chi+\varkappa_n)=\Phi_p-\Phi_n. \tag{72.29}$$

Thus, should we consider the contact between n- and p-type semiconductors, we would arrive at the conclusion that *for electrons the height of the potential barrier separating the two regions is equal to the difference between the respective work functions, the latter coinciding with external contact potential difference. The space charge region separating the n- and p-regions of the semiconductor is termed p-n or n-p junction*. The width of the potential barrier is related to its height. We may write from (72.28)

$$N_dt_n=\sqrt{\frac{2\varepsilon\varepsilon_0N_aN_dU^c}{e^2(N_a+N_d)}}=N_at_p, \tag{72.30}$$

$$t=\sqrt{\frac{2\varepsilon\varepsilon_0U^c}{e^2}\frac{N_a+N_d}{N_aN_d}}. \tag{72.31}$$

The potential barrier U^c acts in a similar way both on the electrons and holes: *the electric field of the p-n junction enhances the transport of minority carriers and prevents the transport of majority carriers from each region*. Neglecting a number of terms in (72.12-13) we obtained a parabolic function (72.23-24). Should these terms be taken into account, a different $V(x)$ would be obtained, but the main result relating to the creation of a space charge and of a built-in field in the region of impurity type inversion would be valid.

By analogy with the p-n junction the terms n-n^+ and p-p^+ junctions are applied to describe a region where a space charge and an electric field are established due to a sufficiently abrupt change in the concentration of impurity of one type.

The space charges and electric fields are also present in solids with a variable forbidden band width. For example, the three-component compound $Cd_yHg_{1-y}Te$ may be produced with any value of $y: 0 < y < 1$. The forbidden band width of the compound varies, depending on y, from 1.5 eV for $y=1$ to 0 eV for $y=0$. In a single crystal of variable composition there is a charge carrier concentration gradient, i.e. an electric field.

The presence of built-in fields results in deviations from Ohm's law for corresponding sections of the circuit. In other words, *the resistance becomes dependent on the current flowing in the semiconductor making the dependence of the current on the voltage nonlinear*. To begin with, write the expression for the current density

$$\mathbf{j} = \sigma \mathbf{E} = \frac{\mathbf{E}}{\rho}. \tag{72.32}$$

Find the potential difference between two arbitrary points $M_1(\mathbf{r}_1)$ and $M_2(\mathbf{r}_2)$:

$$d\varphi = -(\mathbf{E}\,d\mathbf{r});\quad \varphi(M_2) - \varphi(M_1) = -\int_{M_1}^{M_2} (\mathbf{E}\,d\mathbf{r}) = -j\int_{M_1}^{M_2} \rho(\mathbf{r})\left(\frac{\mathbf{j}}{j}\,d\mathbf{r}\right) \tag{72.33}$$

or

$$j = \frac{\varphi(M_1) - \varphi(M_2)}{\int_{M_1}^{M_2} \rho(\mathbf{r})\,dl} = \frac{\varphi_{M_2M_1}}{R_{M_2M_1}}. \tag{72.34}$$

For a sample of constant cross section the current density also will be constant. *The drop in the external voltage $\varphi_{M_2M_1} = jR_{M_2M_1}$ occurs mainly in the section of high resistivity. However, the resistivity of the circuit section M_1M_2 itself depends on the voltage*, therefore one may write

$$j = \frac{\int_{M_1}^{M_2} d\varphi(M)}{\int_{M_1}^{M_2} \dfrac{dl}{e_n\mu_n n_0(M)\, e^{\frac{-e_n\varphi(M)}{kT}} + e_p\mu_p p_0(M)\, e^{\frac{-e_p\varphi(M)}{kT}}}}. \tag{72.35}$$

This means that *the electric conductivity* $\rho(M)$ *is determined not by the initial concentrations $n_0(M)$ and $p_0(M)$ of electrons and holes at each point M, but by the concentrations established as a result of the current flowing in the semiconductor*. The external field deflects the energy bands displacing them by an amount $V(M) = e_n\varphi(M)$ with the resultant change in electron and hole concentrations. This phenomenon allows of a somewhat different explanation. The flooding of the semiconductor by the carriers discussed in Sec. 66 leads to variation of the minority and majority carrier concentrations so that these concentrations become non-equilibrium. It should be noted that if one field direction increases the concentration of the minority carriers, the reversal of the field will result in a decrease in concentration. We saw in Sec. 66 that this is accompanied by the change in majority carrier concentration and, consequently, in resistance. Hence, *the resistance of a section of the circuit depends not only on the magnitude of the current, but on its direction as well.*

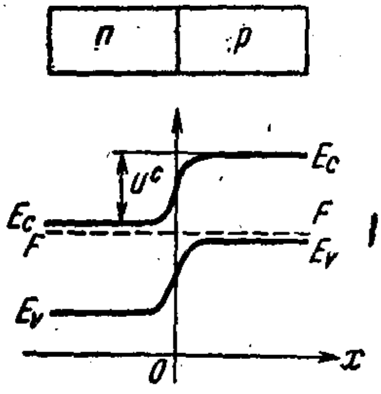


Fig. 108. The formation of a potential barrier for majority carriers on a *p-n* junction

It follows that the resistance of a semiconductor with built-in fields and space charges measured for opposite current directions will be different. The most striking manifestation of this phenomenon is the case of a *p-n* junction. An external voltage applied to a *p-n* junction drops according to (72.33) almost *entirely on the space charge region*, i. e. *on the p-n junction*. Let the external field be directed against the contact field $\mathbf{E}_c$ (from the *p*- to the *n*-region). The external electric field lowers the potential barrier U^c with the result that the electron flow from the *n*-region (or the hole flow from the *p*-region) increases almost exponentially with the applied voltage. When the direction of the applied voltage is reversed, the electric field raises the potential barrier with the result that majority carriers are drawn away from the junction region, and the width of the space charge layer is increased. The current through the junction consists only of minority carriers whose concentration is small, and for this reason the current termed reverse current is also small.

Summary of Sec. 72

1. An inhomogeneous semiconductor contains built-in fields and space charges. The resistance of an inhomogeneous semiconductor depends on the magnitude and the direction of the current flowing in it, since due to the action of the electric field on the minority carriers, injection, extraction, exclusion or accumulation will take place in all sections where the inhomogeneities are localized.

2. The region of abrupt inversion of the type conductivity is termed *p-n* junction. High built-in fields and space charges are a feature of the *p-n* junction. The energy bands are deflected, as shown in Fig. 108. The height of the potential barrier is equal to U^c.

Chapter VIII

OPTICAL AND PHOTOELECTRICAL PHENOMENA IN SEMICONDUCTORS

73. LIGHT-ABSORPTION SPECTRUM

Light, as it penetrates a solid, interacts with it exchanging energy in the course of this interaction. Denote *the light intensity, i. e. the amount of luminous energy crossing a unit cross section normal to the direction of light propagation, by J*. The fraction of the energy reflected on the boundary of the body is determined by the reflection factor, or reflectivity R

$$R = \frac{J_R}{J_0}. \tag{73.1}$$

The dependence of the reflectivity on the frequency R (ω) *or on the wavelength R* (λ) *is termed reflection spectrum.* Light that has penetrated a solid is absorbed by it in compliance with the Bouguer-Lambert law

$$J(x) = J_0(1-R)e^{-\alpha x}, \tag{73.2}$$

where x is the distance from the body boundary to the point in question, measured along the light ray; α is the absorption coefficient. The quantity α^{-1} is equal to the distance at which light intensity is attenuated e times. *The dependence of the absorption coefficient on the frequency* $\alpha(\omega)$ *or on the wavelength* $\alpha(\lambda)$ *is termed absorption spectrum of the body.* Sometimes the place of the absorption coefficient is taken by the *absorption index* $n\varkappa$ related to the absorption coefficient by the expression

$$\alpha = \frac{4\pi n\varkappa}{\lambda} = 4\pi\tilde{\nu}n\varkappa \; (\tilde{\nu} = \lambda^{-1}). \tag{73.3}$$

It follows from the electromagnetic theory of light that the reflectivity for normal incidence may be expressed in terms of refraction and absorption indices as follows:

$$R = \frac{(n-1)^2 + n^2\varkappa^2}{(n+1)^2 + n^2\varkappa^2} = \frac{(n-1)^2 + \frac{\lambda^2}{16\pi^2}\alpha^2}{(n+1)^2 + \frac{\lambda^2}{16\pi^2}\alpha^2}. \tag{73.4}$$

An important feature of the expression (73.4) is the dependence of the reflectivity R on the absorption coefficient: *the reflectivity increases with the increase in the absorption coefficient*, $R \cong 1$ for $\varkappa \gg 1$, i. e. the incident light is almost completely reflected. This explains good light reflection by metals (metallic lustre). Hence, *if a substance absorbs light intensively in some spectral interval it will also effectively reflect it in the same spectral interval*. However, it follows from (73.4) that reflection takes place in the absence of absorption as well: $R \neq 0$ for $\alpha = 0$ and

$$R = \frac{(n-1)^2}{(n+1)^2}. \tag{73.5}$$

For such transparent dielectric as, for instance, glass $n = 1.5$ and $R = 0.04 = 4$ per cent. The values of n for the majority of semiconductors are greater. For example, for germanium $n \cong 4$, therefore the purely dielectric reflectivity for germanium is $R = 0.36 = 36$ per cent.

Table 25 shows the values of the refraction number n for some semiconductors which may be used to assess the value of reflectivity in the region of weak absorption.

Table 25

Substance	n	Substance	n
C (diamond)	2.417	InSb	3.988
Si	3.446	GaP	2.97
Ge	4.006	GaAs	3.348
InP	3.37	GaSb	3.748
InAs	3.428	AlSb	3.188

The Bouguer-Lambert law may be derived from quite general physical principles. Consider a layer $(x, x+dx)$ through which light passes. The amount of energy absorbed in the layer dx wide should be proportional to the layer width and to the luminous energy reaching the layer $x, x+dx, J(x)$. Denoting the proportionality factor between the absorbed energy and the incident energy by α we may write

$$-dJ(x) = \alpha J(x)\, dx. \tag{73.6}$$

Since the absorption of energy decreases the light intensity, there should be a minus in front of dJ ($dJ < 0$). Hence, α *is the amount of energy absorbed from a beam of unit intensity in a layer of unit width*. The equation (73.6) may be easily integrated:

$$J(x) = J(0)\, e^{-\alpha x}. \tag{73.7}$$

The meaning of the quantity α may be made clearer with the aid of a principle repeatedly used before. To this end express the intensity J in terms of the number of photons making up the light beam. If q_1 is the number of photons in a unit volume of the beam, then q_1c photons will pass through a unit area per unit time transporting the energy $q_1c\hbar\omega = q\hbar\omega$, where $q = q_1c$ is the photon flux and, therefore,

$$J(x) = \hbar\omega q(x). \qquad (73.8)$$

The decrease in intensity J means that the number of photons in the beam decreases. The attenuation of the beam may be due either to the scattering of the photons or to their absorption. Denote the probability for a single-photon flux to be absorbed by a single absorption centre by σ and the concentration of such centres by N. A layer dx wide contains $N \cdot dx$ absorption centres. The number of photons absorbed per unit time will be

$$-dq = \sigma q(x)\ N\ dx. \qquad (73.9)$$

Integrating the equation (73.9) we obtain

$$q(x) = q(0)\,e^{-\sigma Nx}. \qquad (73.10)$$

Multiply the equation (73.10) by the photon energy $\hbar\omega$ to get

$$\hbar\omega q(x) = J(x) = \hbar\omega q(0)\,e^{-\sigma Nx} = J(0)\,e^{-\sigma Nx}. \qquad (73.11)$$

Evidently, the relation (73.11) is the Bouguer-Lambert law with the absorption coefficient related to the concentration of absorption centres and to the effective absorption cross section for one photon per unit time:

$$\alpha = \sigma N. \qquad (73.12)$$

The relations of the type (73.12) were repeatedly used previously. The quantity $(\sigma N)^{-1}$ *may be termed mean free path* l_{ph} *of a photon in an absorbing medium:*

$$l_{ph} = (\sigma N)^{-1} = \alpha^{-1}. \qquad (73.13)$$

The quantity α—the absorption coefficient—is *the probability of photon absorption over a unit path.* The effective cross section depends on the photon energy and on the nature of the absorption centres. For a semiconductor containing N_i absorption centres of different nature each with its own effective cross section

$$\alpha_i(\omega) = \sigma_i(\omega)\,N_i. \qquad (73.14)$$

The combined absorption coefficient α is the sum of partial absorption coefficients (the probabilities of independent processes should be added up):

$$\alpha = \sum_i \alpha_i(\omega) = \sum_i \sigma_i(\omega) N_i = \alpha(\omega). \qquad (73.15)$$

Thus, the combined absorption spectrum is made up of the spectra of various absorption centres.

To calculate α (ω) the laws of energy and momentum conservation should be taken into account. However, we may assess α without calculating these quantities. A rough evaluation of σ_i may be based on the following simple considerations. The effective cross section of photon absorption by a matrix atom (or a defect) may be assumed to be equal to the area of the geometrical cross section of the atom (or of the defect) for the frequencies at which the energy conservation law remains valid. Put $\sigma \cong (10^{-16}\text{-}10^{-17})$ cm^2.

For absorption by the matrix atoms $N \cong 10^{22}$ cm^{-3} and $\alpha \cong \cong (10^{-17}\text{-}10^{-16}) \cdot 10^{22}$ cm$^{-1} = (10^5\text{-}10^6)$ cm^{-1}. If we accept this value as the inverse mean free path of photons capable of detaching an electron from the matrix atom, i.e. of photons with energy not less than the width of the forbidden band, $\hbar\omega \geqslant \Delta E_0$, we shall obtain $l_{ph} \cong (10^{-5}\text{-}10^{-6})$ cm $= (0.10\text{-}0.01)$ μm. Despite the inaccuracies involved we obtained a correct estimate of the order of magnitude of the coefficient of light absorption by the matrix atoms. This absorption is termed intrinsic, or fundamental.

The ratio of light absorption by defects (vacancies, impurity atoms) to the fundamental absorption should be equal to the ratio of the defect concentration N_{def} to that of the matrix atoms. For $N_{def} \cong 10^{16}$ cm^{-3} and $\sigma = 10^{-16}$ cm^2 $\alpha = 1$ cm^{-1}; for $N_{def} \cong \cong 10^{18}$ cm^{-3} $\alpha \cong 10^2$ cm^{-1}.

In conclusion let us consider the main types of light absorption in semiconductors.

1. **Intrinsic, or fundamental light absorption** results in the transition of electrons from the bound to the free state, i.e. from the valence to the conduction band. Intrinsic absorption is possible if $\hbar\omega \geqslant \Delta E_0$. Depending on the forbidden band width it takes place in the visible or in the near infrared part of the spectrum.

2. **Impurity absorption** is due to the ionization of the impurity atoms, i.e. to the transition of electrons from impurity atoms to the conduction band or from the valence band to the impurity level.

3. **The absorption by free charge carriers** takes place as a result of their motion induced by the electric field of the light wave. The wave spends some of its energy to accelerate the carriers and is thereby attenuated.

4. The light wave interacts with the lattice vibrations with the resultant change in the number of optical photons. This absorption is termed **lattice absorption**.

5. Light absorption accompanied by the formation of bound electron-hole pairs is termed **exciton absorption.**

6. **Internal-band absorption** is observed in solids with a complicated band pattern, such as the valence band of germanium and silicon.

7. Light absorption by an ensemble of free electrons and holes is termed **plasma absorption.**

It follows from this list that the absorption spectrum should be sensitive to all external influences capable of affecting the state of the matrix atoms, of the defects and of the lattice vibrations. Therefore the factors affecting the absorption spectrum should be expected to include temperature, doping level, pressure, magnetic and electric fields, and corpuscular radiation.

Summary of Sec. 73

1. The attenuation of light which has covered a distance x in a substance is described by the Bouguer-Lambert law:

$$J(x) = J(0)\, e^{-\alpha x}. \tag{73.1s}$$

The absorption coefficient is determined by the effective photon absorption cross section $\sigma(\omega)$ and by the concentration of absorption centres N:

$$\alpha(\omega) = \sigma(\omega)\, N; \quad l_{ph} = (\sigma N)^{-1} = \alpha^{-1}. \tag{73.2s}$$

2. The dependence of the absorption coefficient on the frequency ω or on the wavelength λ is termed absorption spectrum. The dependence of the reflectivity on ω or λ is termed reflection spectrum.

3. Several absorption mechanisms are active in semiconductors: intrinsic, or fundamental absorption; free carrier absorption; internal-band absorption; impurity absorption; lattice absorption; exciton absorption; plasma absorption.

74. LIGHT ABSORPTION BY FREE CHARGE CARRIERS

Consider as the first absorption mechanism the absorption of light by free charge carriers since this mechanism may be described with the aid of classical electrodynamics. The light wave which penetrates a transparent medium acts upon the electrons of the conduction band and upon the holes of the valence band. The electrons are accelerated and thereby increase their energy at the expense of the wave energy. Their excess energy is expended in collisions with the lattice. In the long run the energy of the light wave is converted into the thermal energy of the lattice.

Write the Maxwell equation (in the Gauss system) for the case when extraneous fields and space charges are absent:

$$\text{rot}\,\mathbf{H}=\frac{1}{c}\frac{\partial\mathbf{D}}{\partial t}+\frac{4\pi}{c}\mathbf{j}; \tag{74.1}$$

$$\text{rot}\,\mathbf{E}=-\frac{1}{c}\frac{\partial\mathbf{B}}{\partial t}; \tag{74.2}$$

$$\text{div}\,\mathbf{D}=0; \tag{74.3}$$

$$\text{div}\,\mathbf{B}=0; \tag{74.4}$$

$$\mathbf{D}=\varepsilon\mathbf{E}; \tag{74.5}$$

$$\mathbf{B}=\mu\mathbf{H}; \tag{74.6}$$

$$\mathbf{j}=\sigma\mathbf{E}. \tag{74.7}$$

Consider the wave equation and to this end find the curl of the equation (74.1) or (74.2):

$$\text{rot}\,\text{rot}\,\mathbf{H}=\text{grad}\,\text{div}\,\mathbf{H}-\nabla^2\mathbf{H}=-\nabla^2\mathbf{H}=\frac{1}{c}\,\text{rot}\,\frac{\varepsilon\,\partial\mathbf{E}}{\partial t}+\frac{4\pi}{c}\,\text{rot}\,\sigma\mathbf{E}=$$
$$=\frac{1}{c}\left(\varepsilon\frac{\partial}{\partial t}+4\pi\sigma\right)\text{rot}\,\mathbf{E}=-\frac{1}{c^2}\left(\varepsilon\frac{\partial}{\partial t}+4\pi\sigma\right)\mu\frac{\partial\mathbf{H}}{\partial t}, \tag{74.8}$$

or

$$\nabla^2\mathbf{H}-\frac{\varepsilon\mu}{c^2}\frac{\partial^2\mathbf{H}}{\partial t^2}-\frac{4\pi\mu\sigma}{c^2}\frac{\partial\mathbf{H}}{\partial t}=0. \tag{74.9}$$

To derive the equation (74.9) we made use of the relations (74.1-7). For $\sigma=0$ the equation (74.9) reduces to the wave equation

$$\Delta\mathbf{H}-\frac{\varepsilon\mu}{c^2}\frac{\partial^2\mathbf{H}}{\partial t^2}=0, \tag{74.10}$$

whose solution is a plane wave

$$\mathbf{H}=\mathbf{H}_0\left(t-\frac{(\mathbf{k}^0\mathbf{r})}{v}\right), \tag{74.11}$$

travelling with the velocity $v=\frac{c}{\sqrt{\varepsilon\mu}}$ in the direction of the unit vector $\mathbf{k}^0$. Computing the curl of (74.2) we obtain an equation for $\mathbf{E}$:

$$\Delta\mathbf{E}-\frac{\varepsilon\mu}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2}-\frac{4\pi\mu\sigma}{c^2}\frac{\partial\mathbf{E}}{\partial t}=0. \tag{74.12}$$

The equations (74.12) and (74.9) are conveniently solved by making use of the monochromatic plane wave method. To this end represent $\mathbf{E}=\mathbf{E}(\mathbf{r},\,t)$ in the form of a Fourier integral over all possible frequencies:

$$\mathbf{E}(\mathbf{r},\,t)=\int_{-\infty}^{\infty}\mathbf{E}(\mathbf{r},\,\omega)\,e^{i\omega t}d\omega. \tag{74.13}$$

where the Fourier component of the field $\mathbf{E}(\mathbf{r}, \omega)$ is equal to

$$\mathbf{E}(\mathbf{r}, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{E}(\mathbf{r}, t)\, e^{-i\omega t}\, dt. \tag{74.14}$$

Find the equations for the Fourier components of the field $\mathbf{E}$:

$$\nabla^2 \mathbf{E}(\mathbf{r}, t) = \int_{-\infty}^{\infty} \nabla^2 \mathbf{E}(\mathbf{r}, \omega)\, e^{i\omega t}\, d\omega, \tag{74.15}$$

$$\frac{\partial \mathbf{E}(\mathbf{r}, t)}{\partial t} = \int_{-\infty}^{\infty} i\omega \mathbf{E}(\mathbf{r}, \omega)\, e^{i\omega t}\, d\omega, \tag{74.16}$$

$$\frac{\partial^2 \mathbf{E}(\mathbf{r}, t)}{\partial t^2} = -\int_{-\infty}^{\infty} \omega^2 \mathbf{E}(\mathbf{r}, \omega)\, e^{i\omega t}\, d\omega. \tag{74.17}$$

Substituting (74.15-17) into (74.12) we obtain

$$\int_{-\infty}^{\infty} \left\{ \nabla^2 \mathbf{E}(\mathbf{r}, \omega) + \left[\frac{\omega^2 \varepsilon\mu}{c^2} - i\, \frac{4\pi\mu\sigma\omega}{c^2} \right] \mathbf{E}(\mathbf{r}, \omega) \right\} e^{i\omega t}\, d\omega = 0. \tag{74.18}$$

It follows from (74.18) that the amplitude $\mathbf{E}(\mathbf{r}, \omega)$ of the monochromatic wave satisfies the equation

$$\Delta \mathbf{E}(\mathbf{r}, \omega) + \left[\frac{\omega^2 \varepsilon\mu}{c^2} - i\, \frac{4\pi\mu\sigma\omega}{c^2} \right] \mathbf{E}(\mathbf{r}, \omega) = 0. \tag{74.19}$$

Write the coefficient in front of $\mathbf{E}(\mathbf{r}, \omega)$ in the following form:

$$\frac{\omega^2 \varepsilon\mu}{c^2} - i\, \frac{4\pi\mu\sigma\omega}{c^2} = \frac{\omega^2}{c^2}\, \mathrm{n}^2 (1 - i\varkappa)^2, \tag{74.19'}$$

where n and $\varkappa$ are two unknowns.

Raising the right-hand side of (74.19') to a second power and equating the corresponding real and imaginary terms we obtain

$$\frac{\omega^2 \varepsilon\mu}{c^2} = \frac{\omega^2}{c^2}\, \mathrm{n}^2 (1 - \varkappa)^2 \text{ or } \mathrm{n}^2 (1 - \varkappa^2) = \varepsilon\mu; \tag{74.20}$$

$$i\, \frac{4\pi\mu\sigma\omega}{c^2} = i\, \frac{2\varkappa \mathrm{n}^2 \omega^2}{c^2} \text{ or } \mathrm{n}^2 \varkappa = \frac{2\pi\mu\sigma}{\omega}. \tag{74.21}$$

n and $\varkappa$ may be expressed from (74.21):

$$\mathrm{n} = \sqrt{\frac{\varepsilon\mu}{2} \left(\sqrt{1 + \frac{16\pi^2\sigma^2}{\omega^2\varepsilon^2}} + 1 \right)}, \tag{74.22}$$

$$\mathrm{n}\varkappa = \sqrt{\frac{\varepsilon\mu}{2} \left(\sqrt{1 + \frac{16\pi^2\sigma^2}{\omega^2\varepsilon^2}} - 1 \right)}. \tag{74.23}$$

n bears the name of the *refraction index*, $n\varkappa$—of *the absorption index*. For $\sigma=0$ $n=\sqrt{\varepsilon\mu}$ and $n\varkappa=0$. Making use of (74.19′) we obtain the equation (74.18) in the following form:

$$\Delta \mathbf{E}(\mathbf{r},\omega)+\frac{\omega^2 n^2(1-i\varkappa)^2}{c^2}\mathbf{E}(\mathbf{r},\omega)=0. \qquad (74.24)$$

This equation is easily solved, the general solution being

$$\mathbf{E}(\mathbf{r},\omega)=\mathbf{E}_{0\omega}e^{\pm i\frac{\omega n}{c}(1-i\varkappa)(\mathbf{k}^0\mathbf{r})}=\mathbf{E}_{0\omega}e^{\pm\frac{\omega n\varkappa}{c}(\mathbf{k}^0\mathbf{r})}e^{\pm i\frac{\omega n}{c}(\mathbf{k}^0\mathbf{r})}=\mathbf{E}_{\omega}. \qquad (74.25)$$

Introducing the wave vector $\mathbf{k}=\frac{\omega}{c}\mathbf{k}^0$ write the equation (74.25) for the monochromatic wave retaining only the minus sign:

$$\mathbf{E}_{\omega}(\mathbf{r},t)=\mathbf{E}_{0\omega}e^{-n\varkappa(\mathbf{kr})}\cdot e^{-i[\omega t-n(\mathbf{kr})]}. \qquad (74.26)$$

The expression (74.26) shows that the amplitude of a plane monochromatic wave as it travels through a medium with a finite conductivity σ decreases (if $\varkappa<0$, the medium amplifies the radiation). Since intensity is proportional to the square of the amplitude, we obtain for the intensity measured along the beam $(\mathbf{kr})=kx$

$$J(x)=J_{0\omega}e^{-2n\varkappa kx}=J_{0\omega}e^{-\alpha x}, \qquad (74.27)$$

$$\alpha=2n\varkappa k=\frac{2n\varkappa\omega}{c}=\frac{2n\varkappa\cdot 2\pi}{\lambda}=\frac{4\pi}{\lambda}n\varkappa. \qquad (74.28)$$

Thus, *according to Maxwell's theory, the light absorption coefficient is determined by the conductivity of the substance:*

$$\alpha=\frac{4\pi}{\lambda}n\varkappa=\frac{4\pi}{\lambda}\sqrt{\frac{\varepsilon\mu}{2}\left[\sqrt{1+\frac{16\pi^2\sigma^2}{\omega^2\varepsilon^2}}-1\right]}. \qquad (74.29)$$

Let us assess α for certain value of ε, μ, σ, ω. For that put $\omega\cong 10^{15}\ \mathrm{s}^{-1}$, $\sigma=1\ \mathrm{ohm}^{-1}\,\mathrm{cm}^{-1}=9\times10^{11}\ \mathrm{CGSE}_{\sigma}$, $\varepsilon\cong 10$, $\mu\cong 1$. In this case

$$\frac{16\pi^2\sigma^2}{\omega^2\varepsilon^2}\cong\frac{16\times10\times8\times10^{23}}{10^{30}\times10^2}=1.28\times10^{-6}\ll 1,$$

the inequality being valid up to $\sigma\cong 10^2\ \mathrm{ohm}^{-1}\mathrm{cm}^{-1}$ which is characteristic of heavily doped semiconductors. Expanding the root in (74.29) into a series we obtain

$$\alpha=\frac{4\pi}{\lambda}\sqrt{\frac{\varepsilon\mu}{2}\cdot\frac{1}{2}\cdot\frac{16\pi^2\sigma^2}{\omega^2\varepsilon^2}}=\frac{8\pi^2\sqrt{\varepsilon\mu}\sigma}{\varepsilon\lambda\omega}=\frac{4\pi\sigma\sqrt{\varepsilon\mu}}{\varepsilon c}. \qquad (74.30)$$

Taking into account the estimate obtained for the expression under the radical sign for α or $n\varkappa$ we may reduce the expression

(74.22) to the form

$$n=\sqrt{\varepsilon\mu}. \tag{74.31}$$

For visible light frequency μ is practically unity. Therefore, we may write $n=\sqrt{\varepsilon}$ and

$$\alpha=\frac{4\pi\sigma}{cn}. \tag{74.32}$$

For σ expressed in $\text{ohm}^{-1}\text{cm}^{-1}$ and α in cm^{-1}

$$\alpha\,(\text{cm}^{-1})=\frac{120\pi}{n}\sigma\ (\text{ohm}^{-1}\text{cm}^{-1}). \tag{74.33}$$

If we take $\sigma=1\ \text{ohm}^{-1}\text{cm}^{-1}$ and $n=3.14$, $\alpha=120\ \text{cm}^{-1}$.

Since the conductivity σ is proportional to the charge carrier concentration, the absorption coefficient will be proportional to it, too. If we denote the free carrier concentration by p (to retain the notation n for the refraction index), we may write

$$\alpha=\frac{4\pi\sigma}{cn}=\frac{4\pi}{cn}\frac{e^2\langle\tau\rangle}{m^*}p. \tag{74.34}$$

Thus, *the coefficient of absorption of light by free carriers of concentration p depends on their effective mass, averaged relaxation time, and on the refraction index of the medium.* But the refraction index is actually a function of the wavelength and by force of this *the coefficient of absorption by free charge carriers should also be dependent on the wavelength.* For example, Cauchy long before Maxwell found that the refraction index may be represented in the form

$$n=a+\frac{b}{\lambda^2}+\frac{c'}{\lambda^3}+\ldots, \tag{74.35}$$

where a, b, c' are constants to be determined experimentally for each substance.

The electron theory enables us to find the dependence of n on λ. Consider with this aim in view the polarization of the substance due to the electric field:

$$\mathbf{D}=\mathbf{E}+4\pi\mathbf{P}=\varepsilon\mathbf{E}. \tag{74.36}$$

If $\mathbf{P}=\chi\mathbf{E}$, $\varepsilon=1+4\pi\chi$. To find ε let us calculate the charge displacement. Suppose there are N_l atoms per unit volume each containing f_l charges of the type e_l. The charge displacement is $\mathbf{r}_l$ and the total electric moment per unit volume (the polarization) is

$$\mathbf{P}=\sum_l N_l e_l f_l \mathbf{r}_l=\chi\mathbf{E}. \tag{74.37}$$

To find the displacement of the charge e write its equation of motion. The charge is acted upon by the quasielastic force which returns it into equilibrium position $\mathbf{f}_{ret} = -\gamma\mathbf{r}$, the force of "friction" which depends on the velocity of e and is responsible for the loss of energy $\mathbf{f}_{fr} = -g\dot{\mathbf{r}}$, and the force exercised by the field $\mathbf{E}$:

$$m^*\ddot{\mathbf{r}} = -\gamma\mathbf{r} - g\mathbf{r} + e\mathbf{E}. \tag{74.38}$$

Denoting $\frac{\gamma}{m^*} = \omega_0^2$; $\frac{g}{m^*} = b$ we obtain

$$\ddot{\mathbf{r}} + b\dot{\mathbf{r}} + \omega_0^2\mathbf{r} = \frac{e}{m^*}\mathbf{E} = \frac{e}{m^*}\mathbf{E}_0 e^{-i\omega t}, \tag{74.39}$$

where $\mathbf{E}_0$ is the amplitude of the wave at the point of the charge location. (74.39) is the equation of forced oscillations. The frequency of such oscillations in a stationary state coincides with the frequency of the force, but the phase is different. Therefore, we put

$$\mathbf{r}(t) = \mathbf{r}_0 e^{-i\omega t - i\varphi}. \tag{74.40}$$

Substitute (74.40) into (74.39) to obtain after cancelling out $e^{-i\omega t}$

$$-\omega^2\mathbf{r}_0 e^{-i\varphi} - i\omega b\mathbf{r}_0 e^{-i\varphi} + \omega_0^2\mathbf{r}_0 e^{-i\varphi} = \frac{e}{m^*}\mathbf{E}_0, \tag{74.41}$$

whence

$$\mathbf{r}_0 = \frac{\frac{e}{m^*}\mathbf{E}_0 e^{i\varphi}}{\omega_0^2 - \omega^2 - i\omega b}. \tag{74.42}$$

Substituting the stationary charge displacement into the expression for the polarization, taking the field $\mathbf{E}_0$ outside the summation sign, and cancelling out $\mathbf{E}_0$ we obtain

$$\varepsilon = 1 + 4\pi\chi = 1 + 4\pi\sum_l N_l e_l f_l \frac{\frac{e_l}{m_l^*} e^{i\varphi_l}}{[\omega_{0l}^2 - \omega^2] - i\omega b_l} = \mathrm{n}^2(1 - i\varkappa)^2. \tag{74.43}$$

Equating the real and imaginary parts of the last two expressions in (74.43) we obtain

$$\mathrm{n}^2(1 - \varkappa^2) = 1 + 4\pi\sum_l N_l \frac{e_l^2}{m_l^*} f_l \frac{[\omega_{0l}^2 - \omega^2]\cos\varphi_l - b_l\omega\sin\varphi_l}{[\omega_{0l}^2 - \omega^2]^2 + b_l^2\omega^2} \tag{74.44}$$

$$2\mathrm{n}^2\varkappa = 1 + 4\pi\sum_l N_l \frac{e_l^2}{m_l^*} f_l \frac{[\omega_{0l}^2 - \omega^2]\sin\varphi_l + b_l\omega\cos\varphi_l}{[\omega_{0l}^2 - \omega^2]^2 + b_l^2\omega^2}. \tag{74.45}$$

Solving the equation system (74.44-45) we obtain the expression for n and n$\varkappa$. We see that they indeed are frequency dependent.

Should the computation of atomic polarization be carried out in compliance with quantum-mechanical procedures the form of the frequency dependence of n and n$\varkappa$ would remain the same. The only thing which changes is the meaning of the quantity f_l: in the classical theory f_l is the number of electrons in an atom in a given state; in quantum mechanics f_l is replaced by a quantity

$$f_{ln} = \frac{2m\,|x_{ln}|^2\,\omega_{ln}}{\hbar} = \frac{2m\omega_{ln}}{e^2\hbar}\,|d_{lm}|^2, \tag{74.46}$$

which is termed oscillator force for the $E_l \to E_n$ transition; $\omega_{ln} = \frac{E_l - E_n}{\hbar}$; d_{ln} is the matrix element of the atomic dipole moment $\mathbf{d} = e\mathbf{r}$ induced by the field. Leaving out the details of the computation write the expression for ε:

$$\varepsilon \cong \mathrm{n}^2(\omega) = 1 + \frac{4\pi N}{m^*} \sum_{ln} w_l \frac{f_{ln}}{\omega_{ln}^2 - \omega^2}, \tag{74.47}$$

where w_l is the probability for the atom to be in the state E_l, $w_l \sim \sim e^{-\frac{E_l}{kT}}$. In the derivation of (74.47) the attenuation of the oscillations has not been taken account of. We see from (74.47) and (74.44) that the frequency dependences of n^2 obtained in classical and in quantum physics coincide.

Since conduction electrons are quasifree, when using the expression for n(ω) one would put $\gamma = 0$ and, consequently, $\omega_0 = 0$. Assume for the sake of simplicity that $\varphi = 0$ and that all the conduction electrons are in identical states. This will enable us to rewrite the expressions (74.44) and (74.45) in the form

$$\mathrm{n}^2(1 - \varkappa^2) = 1 - \frac{4\pi}{m^*} pe^2 \frac{1}{\omega^2 + b^2} = 1 - A = C,$$

$$\left(A = \frac{4\pi\sigma}{\langle\tau\rangle} \frac{1}{\omega^2 + b^2}\right); \tag{74.48}$$

$$2\mathrm{n}^2\varkappa = \frac{4\pi pe^2}{m^*} \frac{b\omega}{\omega^4 + b^2\omega^2} = \frac{1}{2} \frac{b}{\omega} A = B. \tag{74.49}$$

Solving the equation system (74.48-49) we obtain

$$\mathrm{n} = \frac{B}{\sqrt{\frac{C}{2}\left[\sqrt{1 + \frac{4B^2}{C^2}} - 1\right]}}, \tag{74.50}$$

$$\mathrm{n}\varkappa = \sqrt{\frac{C}{2}\left[\sqrt{1 + \frac{4B^2}{C^2}} - 1\right]}. \tag{74.51}$$

Since for high frequencies $A \ll 1$ and $B \ll 1$, it follows that

$$n\varkappa = \sqrt{\frac{C}{2}\left[1+\frac{2B^2}{C^2}+\ldots-1\right]} = \frac{B}{\sqrt{C}} \cong B = \frac{2\pi\sigma b}{\langle\tau\rangle\,\omega}\,\frac{1}{\omega^2+b^2}; \quad (74.52)$$

$$n = \frac{B}{\sqrt{\frac{C}{2}\left[1+\frac{2B^2}{C^2}+\ldots-1\right]}} = \sqrt{C} = \sqrt{1-A} = 1-\frac{A}{2} =$$

$$= 1-\frac{2\pi\sigma}{\langle\tau\rangle}\,\frac{1}{\omega^2+b^2}. \quad (74.53)$$

Now find the absorption coefficient

$$\alpha = \frac{4\pi}{\lambda}\,n\varkappa = \frac{8\pi^2\sigma b}{\langle\tau\rangle\,\lambda\omega}\,\frac{1}{\omega^2+b^2} = \frac{4\pi\sigma b}{c\,\langle\tau\rangle}\,\frac{1}{\omega^2+b^2}. \quad (74.54)$$

Comparing the result obtained in (74.54) with (74.32) we may easily see the difference and the similarity of both: according to the phenomenological theory of Maxwell

$$\alpha = \alpha_{phen} = \frac{4\pi\sigma}{c}\cdot\frac{1}{\sqrt{\varepsilon}}; \quad (74.55)$$

according to the electron theory

$$\alpha = \alpha_{el} = \frac{4\pi\sigma}{c}\cdot\frac{b}{\langle\tau\rangle}\cdot\frac{1}{\omega^2+b^2}. \quad (74.56)$$

From (74.56) and (74.55) we obtain

$$\frac{\alpha_{el}}{\alpha_{phen}} = \frac{b\sqrt{\varepsilon}}{\langle\tau\rangle}\,\frac{1}{\omega^2+b^2}. \quad (74.57)$$

For $\omega^2 \ll b^2$ the absorption coefficient is independent of the frequency and is equal to

$$\alpha_{el} = \frac{4\pi\sigma}{c}\cdot\frac{1}{b\,\langle\tau\rangle}. \quad (74.58)$$

Since for long waves the Lorentz electron theory should coincide with the phenomenological theory of Maxwell or, in other words, it should be possible to replace $\varepsilon(\omega)$, $\mu(\omega)$, and $\sigma(\omega)$ by their static values, this should enable us to evaluate the quantity b:

$$b\,\langle\tau\rangle = \sqrt{\varepsilon}; \quad b = \frac{\sqrt{\varepsilon}}{\langle\tau\rangle}. \quad (74.59)$$

Substituting the expression for b from (74.59) into (74.56) we write

$$\alpha = \frac{4\pi\sigma}{c}\,\frac{\sqrt{\varepsilon}}{\langle\tau\rangle^2}\,\frac{1}{\omega^2+\frac{\varepsilon}{\langle\tau\rangle^2}} = \frac{4\pi\sigma}{c}\,\frac{\sqrt{\varepsilon}}{\varepsilon+\omega^2\langle\tau\rangle^2}. \quad (74.60)$$

Thus, for $\omega^2 \ll b^2 = \frac{\varepsilon}{\langle\tau\rangle^2}$ the coefficient of light absorption by free charge carriers is independent of the frequency (or the wavelength) of the light. It remains constant and equal to

$$\alpha = \alpha_0 = \frac{4\pi\sigma}{c\sqrt{\varepsilon}} = \frac{4\pi e^2 \langle\tau\rangle}{c\sqrt{\varepsilon}\, m^*}\, p. \tag{74.61}$$

Assess the magnitude of the effective cross section of absorption of light by an electron. Put $\langle\tau\rangle \cong 10^{-13}$ s; $\varepsilon = 10$, $m^* = 10^{-27}$ g. The result will be $\frac{\alpha}{p} = 3\times10^{-15}$ cm^2.

For the electron or hole concentration $p = 10^{16}$ cm^{-3}, $\alpha = 30$ cm^{-1}, but for $p = 10^{19}$ cm^{-3} $\alpha = 30\,000$ cm^{-1}.

For $\omega^2 \gg b^2$ the absorption coefficient, according to (74.56), depends on the wavelength:

$$\alpha = \frac{4\pi\sigma}{c}\,\frac{b}{\langle\tau\rangle}\,\frac{1}{\omega^2} = \frac{4\pi\sigma}{c\sqrt{\varepsilon}}\,\frac{\varepsilon}{\langle\tau\rangle^2}\,\frac{1}{\omega^2} = \alpha_0\,\frac{\varepsilon}{\langle\tau\rangle^2}\,\frac{\lambda^2}{4\pi^2}. \tag{74.62}$$

Thus, in the high frequency range $\alpha \sim \lambda^2$ and $\alpha < \alpha_0$ since $\omega^2 \gg \frac{\varepsilon}{\langle\tau\rangle^2}$. Assess the frequency at which the dependence changes from $\alpha \sim \sim \lambda^2$ to $\alpha \sim \lambda^0$. To this end express α in terms of α_0:

$$\alpha = \frac{4\pi\sigma}{c}\,\frac{b}{\langle\tau\rangle}\,\frac{1}{\omega^2+b^2} = \frac{4\pi\sigma}{c}\,\frac{1}{b\langle\tau\rangle}\,\frac{b^2}{\omega^2+b^2} = \alpha_0\,\frac{b^2}{\omega^2+b^2}. \tag{74.63}$$

Defining the transition limit by the condition $\alpha = \frac{\alpha_0}{2}$ we obtain $\omega_{lim}^2 = b^2$ or

$$\omega_{lim} = b = \frac{\sqrt{\varepsilon}}{\langle\tau\rangle}; \tag{74.64}$$

For $\langle\tau\rangle = 10^{-13}$ s $\varepsilon \cong 10$, $\omega_{lim} \cong 3\times10^{13}$ s^{-1} or

$$\lambda_{lim} = \frac{2\pi c}{\omega_{lim}} \cong \frac{6.3\times3\times10^{10}}{3\times10^{13}} = 6.3\times10^{-3}\ \text{cm} = 63\ \mu\text{m}.$$

Expressing $\alpha(\lambda)$ in terms of λ_{lim} and taking into account (74.64) and (74.63) we obtain

$$\alpha = \alpha_0\,\frac{\omega_{lim}^2}{\omega_{lim}^2+\omega^2} = \alpha_0\,\frac{\lambda^2}{\lambda_{lim}^2+\lambda^2}. \tag{74.65}$$

The experimental mobility μ_d may be conveniently used instead of the estimates for α and $\langle\tau\rangle$ since α_0 may be expressed in μ_d:

$$\alpha_0 = \frac{4\pi e}{c\sqrt{\varepsilon}}\,\mu_d p. \tag{74.66}$$

The dependence of the absorption coefficient on the wavelength in the electron theory as distinct from the Maxwell phenomenological theory is quadratic with the coefficient being proportional to the mobility or to the averaged relaxation time. But the absorption of the energy of the light wave by free carriers entails their transition from one energy level to another. Therefore one should expect the dependence of the absorption coefficient on the wavelength to be different, according to which scattering mechanism plays the leading part.

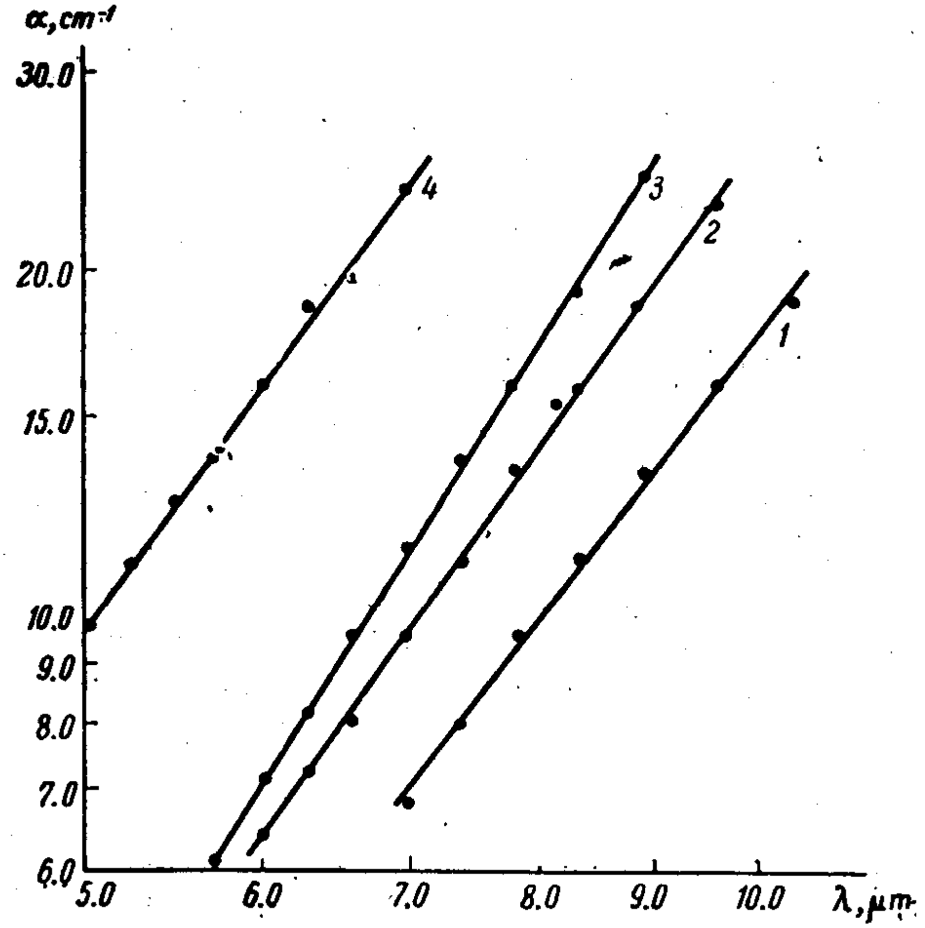


Fig. 109. Free carrier absorption spectrum in cadmium telluride. Electron concentration (cm^{-3}):
$1-1.2\times10^{17}$; $2-1.5\times10^{17}$; $3-2.0\times10^{17}$; $4-3.5\times10^{17}$

When considering the interaction of the free charge carriers with the photons we should keep in mind that a free carrier cannot absorb a photon because the energy and the momentum conservation laws cannot be satisfied at the same time. Light absorption by free carriers takes place via the lattice field. Therefore carrier scattering by crystal defects is important for this process. The absorption coefficient spectral function calculated for different scattering mechanisms will also be different. For one scattering mechanism in case of parabolic bands and a non-degenerate electron gas the expression may be written as follows:

$$\alpha \sim \lambda^{p}, \qquad (74.67)$$

where p assumes the values 3/2, 5/2 and 7/2 for scattering by acoustical vibrations, by optical vibrations, and by impurity ions, respectively. The degeneracy of the electron gas results in some decrease in p, and a deviation from the parabolic band type, i. e. from the quadratic dispersion law, on the other hand—in some increase. For example, calculations carried out for photons of substantial energy $\hbar\omega > \Delta E_0$ show p to be equal to 1.66, 2.83 and 4.21 for scattering by acoustical vibrations, by optical vibrations, and by impurity ions, respectively. Figure 109 shows the spectrum of absorption by electrons in cadmium telluride. It follows from Fig. 109 that the slope of the straight line increases with the electron concentration. This reflects the growing importance of impurity ion scattering. A similar effect takes place as the temperature is lowered.

75. CYCLOTRON RESONANCE

In order to obtain the $\alpha(\lambda)$ dependence we made use of the results of the atomic dispersion theory assuming $\omega_0 = 0$. These results may be obtained by a different method based on the dependence of the conductivity on the frequency of the electric field.

Let $\mathbf{v}$ be the directional velocity of the charge carriers which changes as the result of the application of an external force and of collisions. If τ is the mean time in which the directional velocity is completely lost, the equations of charge carrier motion may be written in the form

$$\frac{d\mathbf{v}}{dt} = \frac{e\mathbf{E}}{m^*} - \frac{\mathbf{v}}{\tau}. \tag{75.1}$$

The first term on the right of (75.1) describes the velocity increment per unit time due to the action of the field $\mathbf{E}$, the second term describes the loss of velocity per unit time due to collisions. Actually we frequently resorted to this equation. Indeed, in a stationary case for a constant field we have

$$\frac{e\mathbf{E}}{m^*} - \frac{\mathbf{v}}{\tau} = 0, \tag{75.2}$$

or

$$\mathbf{v} = \frac{e\tau}{m^*}\mathbf{E} = \mu\mathbf{E}. \tag{75.3}$$

Assume now the external field to be alternating. Since any alternating field may be represented as a superposition of harmonic oscillations, we shall suppose that

$$\mathbf{E} = \mathbf{E}_0 e^{i\omega t} \tag{75.4}$$

and

$$\frac{dv}{dt}=\frac{eE_0}{m^*}e^{i\omega t}-\frac{v}{\tau}. \qquad (75.5)$$

The general solution of the homogeneous equation describes the process of transition to the stationary state. Putting $t \gg \tau$ find *the stationary motion corresponding to the partial solution of the inhomogeneous equation. The stationary motion takes the form of forced oscillations with the frequency of the force.* Therefore put

$$v(t)=v_0 e^{i\omega t}. \qquad (75.6)$$

Substituting (75.6) into (75.5) and cancelling out $e^{i\omega t}$ we obtain

$$\omega v_0=\frac{e}{m^*}E_0-\frac{v_0}{\tau}, \qquad (75.7)$$

or

$$v_0=\frac{eE_0}{m^*\left(i\omega+\frac{1}{\tau}\right)}=\mu\frac{1-i\tau\omega}{1+\omega^2\tau^2}E_0, \qquad (75.8)$$

whence

$$v(t)=\frac{\mu}{1+\omega^2\tau^2}[1-i\tau\omega]E_0e^{i\omega t}=\frac{\mu}{1+\omega^2\tau^2}\left[E-\tau\frac{\partial E}{\partial t}\right]. \qquad (75.9)$$

Multiplying $v(t)$ by the particles' charge and by their concentration p we obtain the current density

$$j=\frac{e^2\tau p}{m^*[1+\omega^2\tau^2]}\left[E-\tau\frac{\partial E}{\partial t}\right]. \qquad (75.10)$$

The current determined by (75.10) consists of two parts. The term proportional to E (or to the real part of v_0) is the ohmic current

$$j_E=\frac{e^2\tau p}{m^*[1+\omega^2\tau^2]}E(t)=\sigma(\omega)E(t), \qquad (75.11)$$

where

$$\sigma(\omega)=\frac{e^2\tau p}{m^*}\frac{1}{1+\omega^2\tau^2}=\frac{\sigma_0}{1+\omega^2\tau^2}. \qquad (75.12)$$

The second part of the current proportional to $\frac{\partial E}{\partial t}$ is *the polarization current*

$$j^{(n)}=-\sigma\tau\frac{\partial E}{\partial t}=\frac{\partial \chi E}{\partial t}, \qquad (75.13)$$

where the polarizability χ is equal to

$$\chi=-\sigma\tau=-\frac{e^2\tau^2 p}{m^*[1+\omega^2\tau^2]}. \qquad (75.14)$$

Using (75.14) we may write for the dielectric permeability

$$\varepsilon(\omega)=1+4\pi\chi=1-\frac{4\pi e^2\tau^2 p}{m^*[1+\omega^2\tau^2]}=1-\frac{4\pi\sigma_0\tau}{1+\omega^2\tau^2}. \qquad (75.15)$$

Putting $\varepsilon=0$ find a critical frequency ω_{cr}

$$\omega_{cr}=\frac{1}{\tau}\sqrt{4\pi\sigma_0\tau-1}. \qquad (75.16)$$

For $\omega>\omega_{cr}$ $\varepsilon>0$ and n is real. The coefficient of light absorption, according to (74.32), assumes the form

$$\alpha=\frac{4\pi\sigma}{cn}=\frac{4\pi}{c\sqrt{1-\frac{4\pi\sigma_0\tau}{\omega^2\tau^2}}}\,\frac{\sigma_0}{1+\omega^2\tau^2}\cong\frac{4\pi\sigma_0}{c\tau^2}\,\frac{1}{\omega^2}, \qquad (75.17)$$

i.e. *in the high frequency range* $\alpha\sim\lambda^2$. For $\omega\ll\omega_{cr}$ $\varepsilon<0$ and n is imaginary, and the absorption is intense:

$$\alpha\cong\frac{4\pi}{\lambda}\sqrt{\frac{4\pi\sigma_0\tau}{1+\omega^2\tau^2}-1}. \qquad (75.18)$$

Thus we arrived at a novel and somewhat unexpected result. In the "low" frequency range the λ dependence of the absorption coefficient is of the $\alpha\sim\lambda^{-1}$ type, the absorption itself being intense. In the high frequency range $\alpha\sim\lambda^2$, and the absorption is comparatively small. Let us assess the critical frequency using (75.16):

$$\omega_{cr}=\frac{1}{\tau}\sqrt{4\pi\sigma_0\tau-1}=\sqrt{\frac{4\pi e^2 p}{m^*}-\frac{1}{\tau^2}}. \qquad (75.19)$$

We see that the critical frequency exists if the carrier concentration is high, i.e.

$$p>\frac{m^*}{\tau^2 4\pi e^2}. \qquad (75.20)$$

Putting $\tau\cong 10^{-13}$ s, $m^*\cong 10^{-27}$ g we obtain

$$p>\frac{10^{-27}}{10^{-26}\times 1.26\times 2.3\times 10^{-18}}\cong 3.5\times 10^{16}\ \text{cm}^{-3}.$$

For lower carrier concentrations the critical frequency is nonexistent. For metals $\omega_{cr}\cong\sqrt{\frac{4\pi e^2 p}{m^*}}\cong 5\times 10^{15}\ \text{s}^{-1}$ ($p\cong 10^{23}\ \text{cm}^{-3}$), i.e. the critical frequency ω_{cr} lies in the ultraviolet region. For $\omega>\omega_{cr}$ α is small. The absorption of light by many metals in the ultraviolet region is indeed small. But for $\omega<\omega_{cr}$ the absorption in metals is very intense. Semiconductors may lack both ω_{cr} and the region of intense "metallic" absorption. Normal absorption is described by the usual relation (75.17) and is proportional to λ^2.

According to (73.4) the reflectivity is determined by the refraction and the absorption coefficients. For this reason the variation of these parameters in the critical frequency regions results in great changes in reflectivity. As the frequency approaches the critical value from the high frequency side reflectivity drops almost to zero after which it increases rapidly (Fig. 110). This phenomenon is termed plasma reflection and may be used to determine the effective mass of free charge carriers. A sufficiently high concentration of free carriers is needed for the effect to be noticeable.

Now consider the absorption of radiant energy by free charge carriers in a semiconductor to which a constant magnetic field of induction **B** has been applied. The equation of carrier motion which *takes account of the collisions and of the electric and magnetic fields* is of the form

$$\frac{\partial \mathbf{v}}{\partial t}=\frac{e\mathbf{E}_0}{m^*}e^{i\omega t}+\frac{e}{m^*}[\mathbf{vB}]-\frac{\mathbf{v}}{\tau}. \qquad (75.21)$$

The solution for the steady state is of the form (75.6). However, the equation for the oscillation amplitude is different from (75.7), namely

$$i\omega\mathbf{v}_0=\frac{e\mathbf{E}_0}{m^*}+\frac{e}{m^*}[\mathbf{v}_0\mathbf{B}]-\frac{\mathbf{v}_0}{\tau}. \qquad (75.22)$$

Re-write the equation (75.22) in the following form:

$$\mathbf{v}_0=\frac{\mu\mathbf{E}_0}{1+i\omega\tau}+\frac{\mu[\mathbf{v}_0\mathbf{B}]}{1+i\omega\tau}. \qquad (75.23)$$

Introducing for the sake of simplicity the notation

$$\frac{\mu\mathbf{E}_0}{1+i\omega\tau}=\mathbf{A};\quad \frac{\mu\mathbf{B}}{1+i\omega\tau}=\boldsymbol{\varphi}, \qquad (75.24)$$

we may write the solution of the vector equation (75.23) in accordance with (37.27) and (37.32) in the form

$$\mathbf{v}_0=\frac{\mathbf{A}+[\mathbf{A}\boldsymbol{\varphi}]+\boldsymbol{\varphi}(\mathbf{A}\boldsymbol{\varphi})}{1+\varphi^2}. \qquad (75.25)$$

We shall confine ourselves to the case of transverse fields since in the case of a longitudinal field the phenomena in solids with scalar carrier effective mass take the same course as in the absence of a magnetic field. Now we can write

$$\mathbf{v}_0=\frac{\mu\mathbf{E}_0}{(1+i\omega\tau)\left[1+\frac{\mu^2B^2}{(1+i\omega\tau)^2}\right]}+\frac{\mu^2[\mathbf{E}_0\mathbf{B}]}{(1+i\omega\tau)^2\left[1+\frac{\mu^2B^2}{(1+i\omega\tau)^2}\right]}. \qquad (75.26)$$

It may be seen from (75.26) that the velocity has a longitudinal $\mathbf{v}_0^{\parallel}$ and a transverse $\mathbf{v}_0^{\perp}$ component (in relation to the electric field).

Energy absorption is due to the fact that the field performs work in displacing the charged particles. It follows from (75.26) that the transverse velocity component does not contribute towards the loss of energy. Therefore we shall consider only the longitudinal component:

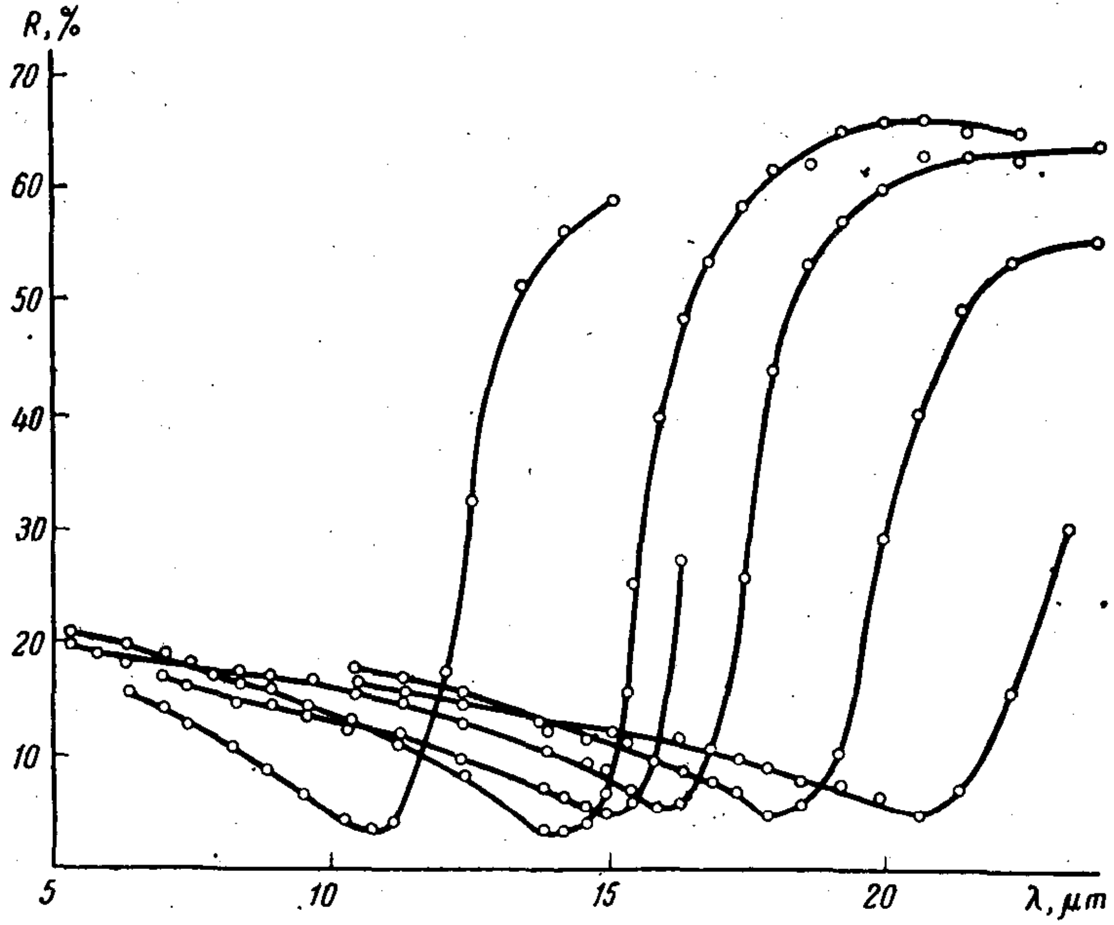


Fig. 110. Plasma reflection in mercury selenide

$$\mathbf{v}_0 = \mu \mathbf{E}_0 \frac{1 + i\omega\tau}{\left[(1 + i\omega\tau)^2 + \frac{e^2\tau^2}{m^{*2}} B^2\right]} =$$

$$= \mu \mathbf{E}_0 \frac{\left[1 + (\omega_c^2 + \omega^2)\tau^2\right] + i\omega\tau\left[(\omega_c^2 - \omega^2)\tau^2 - 1\right]}{\left[1 + (\omega_c^2 - \omega^2)\tau^2\right]^2 + 4\omega^2\tau^2}, \qquad (75.27)$$

$$\omega_c = \frac{eB}{m^*}.$$

As in the case of expression (75.10) the real part of (75.27) is connected with the conductivity current, and the imaginary part — with the polarization current. Using (75.27) we obtain for the conductivity

$$\sigma = e\mu p \frac{1 + (\omega_c^2 + \omega^2)\tau^2}{\left[1 + (\omega_c^2 - \omega^2)\tau^2\right]^2 + 4\omega^2\tau^2}. \qquad (75.28)$$

For $B = 0$ $\omega_c = 0$, and the expression (75.28) reduces to (75.12) which, according tc (74.34), gives for $\alpha(\omega)$ a dependence of the

(74.65) type. Without discussing this case consider the dependence $\alpha(\omega)$ which follows from (74.34) and (75.28) for $B \neq 0$. Denoting

$$\alpha_0 = \frac{4\pi\sigma_0}{cn}, \tag{75.29}$$

we obtain for $\alpha(\omega)$

$$\alpha(\omega) = \alpha_0 \frac{1+(\omega_c^2+\omega^2)\tau^2}{[1+(\omega_c^2-\omega^2)\tau^2]^2+4\omega^2\tau^2}. \tag{75.30}$$

Analyse the expression (75.30) for some values of the constituent parameters.

(1) $\omega_c^2\tau^3 \gg 1$. The expression (75.30) may be simplified as follows: for low frequencies ($\omega \ll \omega_c$)

$$\alpha(\omega) \cong \alpha_0 \frac{\omega_c^2\tau^2}{\omega_c^4\tau^4} = \frac{\alpha_0}{\omega_c^2\tau^2} = \frac{\alpha_0}{\mu^2 B^2}; \tag{75.31}$$

for high frequencies ($\omega \gg \omega_c$)

$$\alpha(\omega) \cong \alpha_0 \frac{\omega^2\tau^2}{\omega^4\tau^4} = \frac{\alpha_0}{\omega^2\tau^2}. \tag{75.32}$$

For the frequencies $\omega \cong \omega_c$ the absorption coefficient is at its maximum:

$$\alpha(\omega) \cong \frac{\alpha_0}{2}. \tag{75.33}$$

It follows that *light absorption is of a "resonance" nature, the resonance being the sharper the better the condition* $\omega_c^2\tau^2 \gg 1$ *is satisfied.* The meaning of this inequality is quite simple: during its mean free time the particle makes several revolutions, the resonance being the sharper the greater the number of revolutions the particle makes. *The phenomenon of resonance absorption of light at the frequency* $\omega \cong \omega_c = \frac{eB}{m^*}$ *is termed cyclotron resonance, and* ω_c*—the cyclotron frequency.*

(2) $\omega_c^2\tau^2 \ll 1$—during its mean free time the particle manages to make less than one revolution.

In the high-frequency range ($\omega^2\tau^2 \gg 1$)

$$\alpha(\omega) = \alpha_0 \frac{\omega^2\tau^2}{\omega^4\tau^4} = \frac{\alpha_0}{\omega^2\tau^2}; \tag{75.34}$$

in the low-frequency range ($\omega^2\tau^2 \ll 1$)

$$\alpha(\omega) \cong \alpha_0. \tag{75.35}$$

Thus, there is no resonance in case of $\omega_c^2\tau^2 \ll 1$. Figure 111 shows the $\alpha(\omega)$ dependence for three different values of $\omega_c\tau = \mu B$. For the investigation of the cyclotron resonance a definite frequency is chosen, and the resonance is attained by changing the magnetic field, i.e. ω_c. After ω_c has been determined the effective mass m^* may be calculated. Cyclotron resonance enables information on the effective mass to be obtained even if it is a tensor quantity. Figure 112 shows the relative values of the absorption peaks in germanium. The cyclotron resonance may be explained as being due to the quantization of electron and hole energy in the magnetic field which results in the creation of the Landau levels. The distance between the levels is

$$\hbar\omega_c = \frac{\hbar e B}{m^*}. \tag{75.36}$$

When the photon energy is equal to the distance between the levels

$$\hbar\omega = \hbar\omega_c, \tag{75.37}$$

photon absorption is possible. But the equation (75.37) is the condition for the cyclotron resonance.

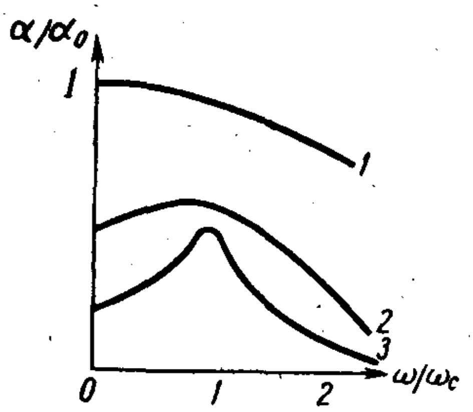


Fig. 111. The conditions for observing cyclotron resonance: $1-\omega_c\tau=0.2$; $2-\omega_c\tau=1$; $3-\omega_c\tau=2$

To illustrate the calculations performed in Secs. 74, 75 let us review some experimental data. Figure 113 shows the dependence of the absorption coefficient of p-silicon in the range from 1 to 2.5 μm. A sharp decrease in α in the range of $\lambda \cong 1.1$ μm is due to the decrease in intrinsic absorption which becomes free carrier absorption (the right-hand side of the dashed line). Eventually the sample under investigation was irradiated with neutrons, its conductivity decreased, and the pattern of the absorption curve was seen to change greatly.

Figure 114 presents in a double logarithmic scale the absorption curves for a tellurium single crystal. The x-axis is graduated in the wave-number logarithms $\bar{\nu} = \lambda^{-1}(\times 10^{3}\ \text{cm}^{-1})$. It is clearly evident that the logarithm of the absorption coefficient in the range from 300 cm^{-1} to 2000 cm^{-1} is related linearly to the logarithm of the wave number. The absorption coefficient, moreover, increases with temperature due to the increase in free carrier concentration.

Figure 115 shows the absorption due to holes injected into n-germanium. The absorption by free carriers may be utilized to study phenomena accompanying non-equilibrium processes taking place on the contacts.

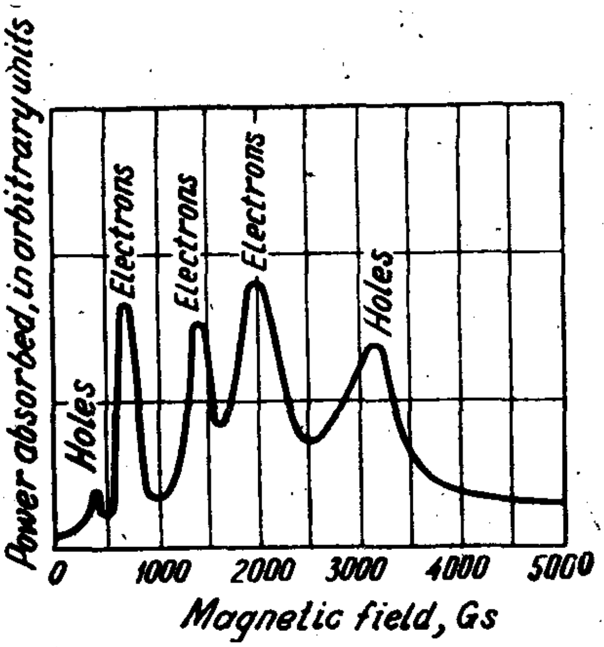


Fig. 112. Absorption peaks of cyclotron resonance in germanium

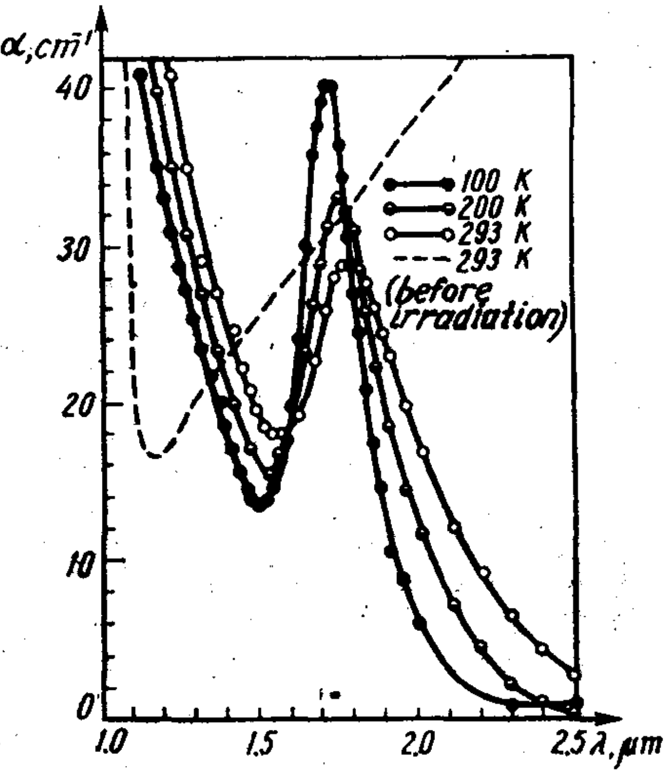


Fig. 113. The effect of neutron irradiation of *p*-silicon on the free-carrier light absorption

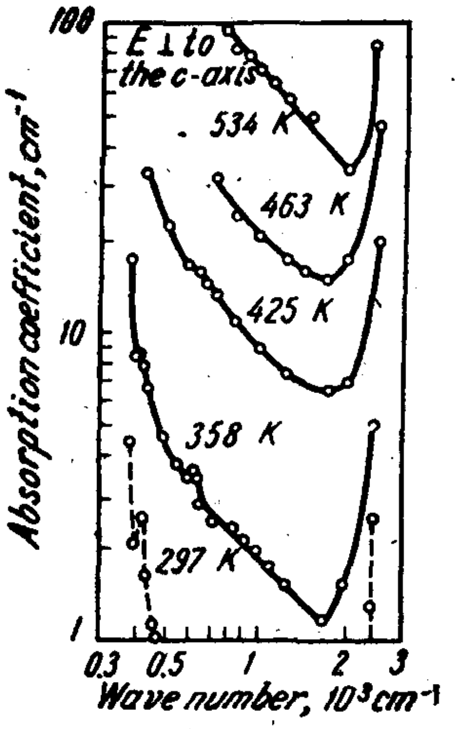


Fig. 114. The temperature dependence of free carrier light absorption in tellurium single crystals

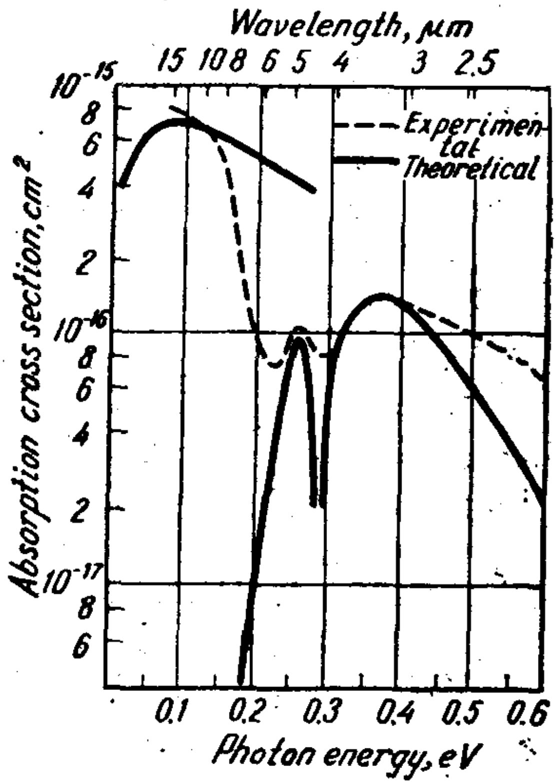


Fig. 115. Light absorption by holes injected in *n*-germanium

Summary of Secs. 74-75

1. The absorption of radiant energy by free charge carriers is due to the fact that the eleotric field of the light wave produces ohmic current and sustains it by its energy, the latter being converted into Joule heat. The absorption coefficient is proportional to the conductivity:

$$\alpha = \frac{4\pi}{cn}\sigma. \tag{75.1s}$$

2. The frequency dependence of the absorption coefficient may be found either with the help of the dispersion theory or on the basis of the frequency dependence of the conductivity. In both cases the frequency dependence of α is of the form

$$\alpha = \frac{\alpha_0 \omega_{lim}^2}{\omega_{lim}^2 + \omega^2}, \tag{75.2s}$$

i.e. in the range $\omega^2 \ll \omega_{lim}^2$ $\alpha = \alpha_0$, and in the range $\omega^2 \gg \omega_{lim}^2$ α decreases as $\alpha \sim \omega^{-2}$. The physical meaning of ω_{lim} in various theories is not the same but in all cases ω_{lim} is related to the relaxation time.

3. The application of a magnetic field to a semiconductor results in a decrease in its conductivity, and because of this the free carrier absorption decreases as well. The expression for $\alpha(\omega)$ assumes the form

$$\alpha(\omega) = \frac{4\pi e\mu p}{cn} \frac{1 + (\omega_c^2 + \omega^2)\tau^2}{[1 + (\omega_c^2 - \omega^2)\tau^2]^2 + 4\omega^2\tau^2}, \tag{75.3s}$$

where p is the carrier concentration, n—the refraction index.

4. When the radiation frequency ω coincides with the cyclotron frequency ω_c the absorption coefficient increases in comparison with the frequencies $\omega \neq \omega_c$. This phenomenon is termed cyclotron, or diamagnetic, resonance. This effect is noticeable when the particle makes at least one revolution during the relaxation time. The better the inequality $\omega_c^2\tau^2 \gg 1$ is satisfied the more pronounced is the resonance. From the point of view of quantum mechanics the cyclotron resonance is the result of the quantization of the energy of the particles in the magnetic field. A photon is absorbed when the distance between the Landau levels is equal to the energy of the photon:

$$\hbar\omega = \hbar\omega_c; \quad \omega = \omega_c. \tag{75.4s}$$

Cyclotron resonance enables m^* to be determined. It is observed at low temperatures, usually in the liquid helium range.

5. In the course of light absorption by free charge carriers the latter interact with the lattice. This fact is responsible for the

dependence of the absorption coefficient on the scattering mechanism. The absorption in non-degenerate semiconductors with standard-type energy band pattern for scattering by acoustical and optical lattice vibrations and by impurity ions is proportional to $\lambda^{3/2}$, $\lambda^{5/2}$, and $\lambda^{7/2}$, respectively. In case of degeneracy or of deviation of the dispersion law from the quadratic the power of the wavelength dependence for different scattering mechanisms increases respectively.

76. INTRINSIC LIGHT ABSORPTION

We shall use the perturbation theory to describe intrinsic light absorption. The energy of the electron in a light wave described by its electric field intensity **E** and magnetic field induction **B** will be used as perturbation. The vector potential **A** (**r**, t) may be conveniently introduced to describe **E** and **B** (in the Gauss system):

$$\mathbf{E}=-\frac{1}{c}\frac{\partial\mathbf{A}}{\partial t}, \tag{76.1}$$

$$\mathbf{B}=\operatorname{rot}\mathbf{A}. \tag{76.2}$$

The Hamiltonian of the electron in a crystal acted upon by radiation in the effective mass approximation takes, according to (23.29) and (23.11), the following form:

$$\hat{\mathbf{H}}=\frac{\left(\mathbf{P}-\frac{e}{c}\mathbf{A}\right)^2}{2m^*}=-\frac{\hbar^2}{2m^*}\Delta+\frac{i\hbar e}{m^*c}(\mathbf{A}\nabla)+\frac{i\hbar e}{2m^*c}\operatorname{div}\mathbf{A}+\frac{e^2\mathbf{A}^2}{2m^*c^2}. \tag{76.3}$$

For weak light intensities which may be obtained from conventional light sources the last term proportional to $\mathbf{A}^2$ may be neglected as compared to the linear term *.

Since the vector potential should satisfy the Lorentz condition

$$\operatorname{div}\mathbf{A}=0, \tag{76.4}$$

we may separate the perturbation operator from (76.3):

$$\hat{\mathbf{W}}=\frac{i\hbar e}{m^*c}(\mathbf{A}\nabla)=-\frac{e}{c}\left(\mathbf{A},\ -\frac{i\hbar\nabla}{m^*}\right)=-\frac{1}{c}(\mathbf{A}\mathbf{j}), \tag{76.5}$$

where **j** is the current density operator

$$\hat{\mathbf{j}}=e\hat{\mathbf{v}}=-\frac{ie\hbar}{m^*}\nabla. \tag{76.6}$$

* The term $\mathbf{A}^2$ should be taken into account when the semiconductors are illuminated with high-intensity light produced by lasers. The result will be a description of two-photon processes.

Hence, the Hamilton operator for the electrons of a semiconductor in a light field is of the form

$$\hat{H} = \hat{H}^0 + \hat{W} = -\frac{\hbar^2}{2m^*}\Delta + \hat{W}. \qquad (76.7)$$

As was demonstrated in Sec. 55, the perturbation results in transitions from state to state. Consider in the valence band a state with the energy $E_1(\mathbf{k}_1)$ and the corresponding wave function $\psi_{1\mathbf{k}_1}(\mathbf{r}, t)$:

$$\psi_{1\mathbf{k}_1}(\mathbf{r}, t) = \frac{1}{L^{3/2}} e^{i(\mathbf{k}_1\mathbf{r})} e^{-i\frac{E_1(\mathbf{k}_1)t}{\hbar}}. \qquad (76.8)$$

Take in the conduction band a state with the energy $E_2(\mathbf{k}_2)$ and a wave function $\psi_{2\mathbf{k}_2}(\mathbf{r}, t)$:

$$\psi_{2\mathbf{k}_2}(\mathbf{r}, t) = \frac{1}{L^{3/2}} e^{i(\mathbf{k}_2\mathbf{r})} e^{-i\frac{E_2(\mathbf{k}_2)t}{\hbar}}. \qquad (76.9)$$

To calculate the transition probability one should define the perturbation $\hat{W} = \hat{W}(\mathbf{r}, t)$. Take the vector potential in the form of a plane wave:

$$\mathbf{A}(\mathbf{r}, t) = \mathbf{A}_0 e^{i[\omega t - (\mathbf{g}\mathbf{r})]}. \qquad (76.10)$$

According to (76.5) and (76.10) we have for the perturbation operator $\hat{W}$:

$$\hat{W} = \frac{ie\hbar}{cm^*}(\mathbf{A}\nabla) = \frac{ie\hbar}{cm^*} e^{i[\omega t - (\mathbf{g}\mathbf{r})]}(\mathbf{A}_0\nabla). \qquad (76.11)$$

Using the wave functions $\psi_{1\mathbf{k}_1}(\mathbf{r}, t)$ and $\psi_{2\mathbf{k}_2}(\mathbf{r}, t)$ calculate the matrix element of $\hat{W}$:

$$\int\limits_{(L^3)} \psi^*_{1\mathbf{k}_1}(\mathbf{r}, t)\, \hat{W}(\mathbf{r}, t)\, \psi_{2\mathbf{k}_2}(\mathbf{r}, t)\, d\tau =$$

$$= \frac{1}{L^3} e^{\frac{i}{\hbar}[E_1(\mathbf{k}_1) - E_2(\mathbf{k}_2) + \hbar\omega]t} \frac{ie\hbar}{cm^*} \int\limits_{L_3} e^{-i(\mathbf{k}_1\mathbf{r})} e^{-i(\mathbf{g}\mathbf{r})} (\mathbf{A}_0\nabla) e^{i(\mathbf{k}_2\mathbf{r})}\, d\tau =$$

$$= e^{\frac{i}{\hbar}[E_1(\mathbf{k}_1) - E_2(\mathbf{k}_2) + \hbar\omega]t} (-1)\frac{e\hbar}{cm^*}(\mathbf{A}\mathbf{k}_2)\frac{1}{L^3}\int e^{-i(\mathbf{k}_1 + \mathbf{g} - \mathbf{k}_2,\, \mathbf{r})}\, d\tau =$$

$$= -\frac{e\hbar}{cm^*}(\mathbf{A}_0\mathbf{k}_2)\, e^{\frac{i}{\hbar}[E_1(\mathbf{k}_1) - E_2(\mathbf{k}_2) + \hbar\omega]t}\, \delta_{\mathbf{k}_1 + \mathbf{g},\, \mathbf{k}_2}. \qquad (76.12)$$

The expression (76.12) shows the matrix element to be non-zero only if

$$\mathbf{k}_1 + \mathbf{g} = \mathbf{k}_2, \qquad (76.13)$$

or

$$\mathbf{P}_2 = \mathbf{P}_1 + \hbar\mathbf{g}, \tag{76.14}$$

i. e. the quasimomentum conservation law should be satisfied in the process of light absorption: *the quasimomentum of the eventual state is equal to the vector sum of the quasimomentum of the initial state and of the photon momentum.*

For $\mathbf{k}_1 = 0$, $\mathbf{k}_2 = \mathbf{g}$. But such transitions are impossible since in this case $(\mathbf{A}_0\mathbf{k}_2) = (\mathbf{A}_0\mathbf{g}) = 0$ (this being the condition for the light waves to be transverse). The quasimomentum of thermal electrons is equal to $\mathrm{P} = \sqrt{2m^*kT}$. For $T = 300$ K and $m^* = 10^{-27}$ g we obtain $\mathrm{P} \cong 10^{-20}$ g·cm/s and $k \cong 10^7$ cm^{-1}. For the light of wavelength $\lambda \cong 1$ µm $\mathrm{g} = \frac{2\pi}{\lambda} \cong 6 \times 10^4$ cm^{-1}, which is much less than k for thermal electrons. In this case we obtain, neglecting $\hbar\mathbf{g}$ as compared to $\mathbf{P}_1$,

$$\mathbf{P}_2 = \mathbf{P}_1; \ \mathbf{k}_2 = \mathbf{k}_1. \tag{76.15}$$

The transitions from the valence band to the conduction band complying to the selection rules (76.15), *i. e. retaining the wave vector of the electron, are termed direct, or vertical.* The electron having absorbed a photon goes over from a point in the Brillouin valence zone to the equivalent point in the Brillouin conduction zone.

It follows from (76.12) that the matrix element of the perturbation operator includes the term $e^{\frac{i}{\hbar}[E_1(\mathbf{k}_1) - E_2(\mathbf{k}_2) + \hbar\omega]t}$. When the transition probability per unit time is calculated this factor yields a δ-function (see Sec. 55)

$$\delta[E_1(\mathbf{k}_1) - E_2(\mathbf{k}_2) + \hbar\omega],$$

which provides for the energy conservation law to be satisfied in the process of light absorption:

$$E_2(\mathbf{k}_2) = E_1(\mathbf{k}_1) + \hbar\omega. \tag{76.16}$$

In accordance with (55.23) the probability of electron transition from a unit volume of the $\mathbf{k}_1$ space into a unit volume of the $\mathbf{k}_2$ space per unit time is equal to

$$w(\mathbf{k}_1, \mathbf{k}_2) = \frac{2\pi}{\hbar}\frac{e^2\hbar^2}{c^2m^{*2}}(\mathbf{A}_0\mathbf{k}_2)^2\,\delta[E_1(\mathbf{k}_1) - E_2(\mathbf{k}_2) + \hbar\omega]\,\delta_{\mathbf{k}_1+\mathbf{g},\,\mathbf{k}_2} \tag{76.17}$$

In future we shall presume the energy and quasimomentum conservation laws to be satisfied and will omit the corresponding δ-functions. Express the electron transition probability (76.17) in terms of the number of photons passing through the semiconductor. To this end make use of the fact that the average density of light energy is $\frac{\varepsilon E_0^2}{8\pi}$ $\left(\text{the instantaneous density is } \frac{\varepsilon E_0^2}{8\pi}+\frac{\mu H^2}{8\pi}= =\frac{\varepsilon E^2}{4\pi}\right)$, and the energy flux is $c\frac{\varepsilon E_0^2}{n8\pi}$, where n is the refraction index of the substance and $\frac{c}{n}$—the velocity of light in it. Find q by dividing the energy flux by the energy of one quantum:

$$q=\frac{c\varepsilon E_0^2}{n8\pi\hbar\omega}. \qquad (76.18)$$

Express the probability of electron transition in terms of the photon flux by expressing A_0^2 in terms of q. First find the connection between E_0 and A_0:

$$\mathbf{E}=-\frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}=-\frac{i\omega}{c}\mathbf{A}_0 e^{i[\omega t-(\mathbf{gr})]}=\frac{\omega \mathbf{A}_0}{c}e^{i\left[\omega t-(\mathbf{gr})-\frac{\pi}{2}\right]} \qquad (76.19)$$

and

$$E_0=\frac{\omega}{c}A_0=gA_0. \qquad (76.20)$$

It follows from (76.18) and (76.20) that

$$q=\frac{c\varepsilon E_0^2}{n8\pi\hbar\omega}=\frac{\varepsilon A_0^2\omega}{8\pi\hbar nc}, \qquad (76.21)$$

and

$$A_0^2=\frac{8\pi\hbar cn}{\varepsilon\omega}q. \qquad (76.22)$$

For the transition probability (76.17) using (76.22) we obtain

$$w(\mathbf{k}_1, \mathbf{k}_2)=\frac{16\pi^2\hbar^2e^2nk_2^2\cos^2\theta}{c\varepsilon m^{*2}\omega}q \quad (\theta=(\widehat{\mathbf{A}_0, \mathbf{k}_2})). \qquad (76.23)$$

The transition of the electron from one state to another is possible only as a result of photon absorption. Therefore $w(\mathbf{k}_1, \mathbf{k}_2)$ is the photon absorption probability. Since it is proportional to the photon flux q, we obtain the effective cross section of absorption of a single photon flux by one electron dividing $w(\mathbf{k}_1, \mathbf{k}_2)$ by the photon flux q:

$$\sigma_q=\frac{w}{q}=\frac{16\pi^2\hbar^2e^2nk_2^2\cos^2\theta}{c\varepsilon m^{*2}\omega} \qquad (76.24)$$

Assess the effective cross section of absorption of a single photon flux by one electron putting $m^*\cong 10^{-27}$ g, $\varepsilon\cong 16$, $n\cong 4$, $k_2\cong \cong 10^8$ cm^{-1}. For $\omega\cong 10^{14}$ s^{-1} and $\cos^2\theta=1/3$; we obtain $\sigma_q=$

$= 10^{-26}\ \mathrm{cm}^2$. This value is equal to the cross section of an electron in classical physics. When calculating σ_q we assumed $\cos^2\theta = 1/3$ which is equal to $\langle\cos^2\theta\rangle$:

$$\langle\cos^2\theta\rangle = \frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi}\cos^2\theta\sin\theta\, d\theta\, d\varphi = \frac{1}{3}. \tag{76.25}$$

The probability (76.23) per photon of the flux $\frac{c}{n}$ is w_{ph}:

$$w_{ph} = \sigma_q\frac{c}{n} = \frac{16\pi^2\hbar^2e^2k_2^2\cos^2\theta}{\varepsilon m^{*2}\omega}. \tag{76.26}$$

To evaluate the absorption probability of one photon by one electron, (76.26) should be divided by the flux produced by an electron, $\frac{\hbar k_2}{m^*}$. Hence, the probability of a photon being absorbed by an electron is

$$w_{c\,ph} = \frac{w_{ph}}{\frac{\hbar k_2}{m^*}} = \frac{16\pi^2\hbar e^2\cos^2\theta k_2}{\varepsilon m^*\omega}. \tag{76.27}$$

Now find the absorption coefficient α. An element of volume $d\tau_{\mathbf{k}_2}$ contains $d\tau_{\mathbf{k}_2}(4\pi^3)^{-1}f_{p_2}(\mathbf{k}_2)$ free states and $d\tau_{\mathbf{k}_2}(4\pi^3)^{-1}f_2(\mathbf{k}_2)$ occupied states. An element of volume $d\tau_{\mathbf{k}_1}$ contains $d\tau_{\mathbf{k}_1}(4\pi^3)^{-1}f_1(\mathbf{k}_1)$ occupied and $d\tau_{\mathbf{k}_1}(4\pi^3)^{-1}f_{p_1}(\mathbf{k}_1)$ free states. Since the probabilities of the direct and reverse transitions are equal, when calculating the light absorption coefficient both direct and reverse light-induced transitions should be taken into account. Spontaneous transitions (recombination transitions) will be neglected. The number of photons absorbed per unit time will be

$$\delta q = \int\int[f_1(\mathbf{k}_1)f_{p2}(\mathbf{k}_2)w(\mathbf{k}_1,\mathbf{k}_2) - f_2(\mathbf{k}_2)f_{p1}(\mathbf{k}_1)w(\mathbf{k}_2,\mathbf{k}_1)]\times \times\frac{d\tau_{\mathbf{k}_1}}{4\pi^3}\frac{d\tau_{\mathbf{k}_2}}{4\pi^3}. \tag{76.28}$$

The first term in (76.28) describes the number of absorbed photons and the second—the number of radiated ones. Under normal conditions the energy level occupancy by electrons coincides with their equilibrium distribution over states.

$$\left.\begin{aligned} f_1(\mathbf{k}_1) &= \frac{1}{e^{\frac{E_1-F}{kT}}+1} \cong 1;\ f_2(E_2) \cong 0; \\ f_{p1}(\mathbf{k}_1) &= \frac{1}{e^{\frac{F-E_1}{kT}}+1} \cong 0;\ f_{p_2}(E_2) \cong 1, \end{aligned}\right\} \tag{76.29}$$

i. e. the valence band is practically filled and the conduction band practically free. Therefore reverse transitions may be neglected. (However, should an inverse energy level occupancy be created, (76.28) would turn negative, and the semiconductor would amplify radiation instead of absorbing it.) Multiplying (76.28) by $\hbar\omega$ we obtain the energy absorbed in a unit of the semiconductor volume per unit time.

But according to (73.6) α is the quantity of energy absorbed in unit volume per unit time:

$$\alpha=\frac{\delta q\hbar\omega}{q\hbar\omega}. \tag{76.30}$$

Integrating (76.28) with (76.23) and (76.17) taken into account we obtain

$$\alpha=\frac{\hbar^2e^2\mathrm{n}}{\pi^4c\varepsilon m^{*2}\omega}\int\limits_{V_{\mathbf{k}_2}}\cos^2\theta\mathbf{k}_2^2\delta\left(E_1(\mathbf{k}_2)-E_2(\mathbf{k}_2)+\hbar\omega\right)d\tau_{\mathbf{k}_2}. \tag{76.31}$$

Exclude the dependence of α on E_1, E_2, and k_2^2 by making use of the fact that $E_2=E_1+\hbar\omega$. It should, moreover, be kept in mind that the effective mass m^* which enters all expressions (76.1-31) is the effective mass of the electron in the valence band or the hole mass. Express k_2^2 in terms of ω:

$$\frac{\hbar^2k_2^2}{2m_n^*}=E_2-E_c=E_1+\hbar\omega-E_c=E_v-\frac{\hbar^2k_1^2}{2m_p^*}+\hbar\omega-E_c. \tag{76.32}$$

For $k_2=k_1$ we obtain from (76.32)

$$\delta\left(E_1(\mathbf{k}_2)-E_2(\mathbf{k}_2)+\hbar\omega\right)=\delta\left(E_v-\frac{\hbar^2k_2^2}{2m_p^*}-E_c-\frac{\hbar^2k_2^2}{2m_n^*}+\hbar\omega\right)=$$

$$=\delta\left((\hbar\omega-\Delta E_0)-\frac{\hbar^2k_2^2}{2m_{red}^*}\right), \tag{76.33}$$

where

$$m_{red}^*=\frac{m_n^*m_p^*}{m_n^*+m_p^*} \tag{76.34}$$

is *the reduced effective mass of the electron and the hole*. Using (76.33) we may write for the absorption coefficient

$$\alpha=\frac{\hbar^2e^2\mathrm{n}}{\pi^4c\varepsilon m_p^{*2}\omega}\cdot\frac{4\pi}{3}\int\limits_0^\infty k_2^4\delta\left(\hbar\omega-\Delta E_0-\frac{\hbar^2k_2^2}{2m_{red}^*}\right)dk_2. \tag{76.35}$$

The factor $\frac{4\pi}{3}$ is due to integration over the angles in which the polar angle is measured from the direction of the vector potential. To calculate the integral in (76.35) put

$$\frac{\hbar^2 k_2^2}{2m^*_{red}} = x; \quad k_2 = \left(\frac{2m^*_{red}}{\hbar^2}\right)^{1/2} x^{1/2}, \tag{76.36}$$

whence

$$k_2^4 dk_2 = \frac{1}{2}\left(\frac{2m^*_{red}}{\hbar^2}\right)^{5/2} x^{3/2}\, dx. \tag{76.37}$$

Now the integral may be easily calculated:

$$\int_0^\infty \frac{1}{2}\left(\frac{2m^*_{red}}{\hbar^2}\right)^{5/2} x^{3/2} \delta\,(\hbar\omega - \Delta E_0 - x)\, dx = \frac{1}{2}\left(\frac{2m^*_{red}}{\hbar^2}\right)^{5/2} (\hbar\omega - \Delta E_0)^{3/2}. \tag{76.38}$$

Substituting (76.38) into (76.35) we obtain

$$\alpha = \frac{\hbar^2 e^2 \mathrm{n}}{\pi^4 c\omega e m_p^{*2}} \cdot \frac{4\pi}{3} \cdot \frac{1}{2}\left(\frac{2m^*_{red}}{\hbar^2}\right)^{5/2} (\hbar\omega - \Delta E_0)^{3/2} =$$

$$= \frac{2}{3\pi^3} \frac{e^2}{c\hbar^3\omega} \frac{(2m^*_{red})^{5/2}}{m_p^{*2}\mathrm{n}} (\hbar\omega - \Delta E_0)^{3/2} = A\,(\hbar\omega - \Delta E_0)^{3/2},$$

$$A = \frac{2}{3\pi^3} \frac{e^2}{c\hbar^3\omega} \frac{(2m^*_{red})^{5/2}}{m_p^{*2}\mathrm{n}}. \tag{76.39}$$

The expression (76.39) shows that for vertical transitions in the sma'l $(\hbar\omega - \Delta E_0)$ range the absorption coefficient is proportional to $(\hbar\omega - \Delta E_0)^{3/2}$. For $\hbar\omega < \Delta E_0$ $\alpha = 0$: there is a well-defined boundary of intrinsic absorption from the low frequency side. The intrinsic absorption boundary is determined by the (optical) forbidden band width for vertical transitions:

$$\omega_b = \frac{\Delta E_0}{\hbar}; \quad \lambda_b = \frac{hc}{\Delta E_0}. \tag{76.40}$$

The forbidden band width in eV may be calculated from the relation

$$\lambda_b\,(\mu\mathrm{m}) = \frac{1.24}{\Delta E_0\,(\mathrm{eV})}. \tag{76.41}$$

It follows from (76.39) that near the intrinsic absorption boundary α is determined by the 3/2 power of the difference between the photon energy and the forbidden band width.

The expression (76.39) is valid only for transitions in semiconductors having spherical energy surfaces and extrema lying at one point, for instance, in the centre of the Brillouin zone. Consider now the periodic crystal field. The electron Hamiltonian which takes account of the Lorentz condition (76.4) may be written in the form:

$$\hat{\mathrm{H}} = -\frac{\hbar^2}{2m_0}\Delta + U(\mathbf{r}) + \frac{i\hbar e}{m_0 c}(\mathbf{A}\nabla) = \hat{\mathrm{H}}_0 + \hat{\mathrm{W}}, \qquad (76.42)$$

where the perturbation operator is

$$\hat{\mathrm{W}} = \frac{i\hbar e}{m_0 c}(\mathbf{A}\nabla) = -\frac{eA_0}{m_0 c} e^{i[\omega t - (\mathbf{gr})]}(\mathbf{a}^0\hat{\mathbf{p}}). \qquad (76.43)$$

In this expression the mass of the free electron m_0 takes the place of the effective mass. Moreover, a unit vector $\mathbf{a}^0$ determining the polarization of the wave is introduced to facilitate the calculations. The Bloch functions $\psi_{n\mathbf{k}} = e^{i(\mathbf{kr})}\varphi_{n\mathbf{k}}(\mathbf{r})$ with two indices showing the position of the point $\mathbf{k}$ in the Brillouin zone and the zone number n are used to calculate the matrix elements of the perturbation. The time dependence of the Bloch function is the same as that of the de Broglie wave, therefore the calculation of the transition probability will again yield a δ-function leading to the energy conservation law for the electron transition from the valence band to the conduction band. Omitting the time dependence we obtain for the matrix element of the perturbation operator

$$W_{n\mathbf{k}_1, n'\mathbf{k}_2} = \int e^{-i(\mathbf{k}_1\mathbf{r})}\varphi^*_{n\mathbf{k}_1}\frac{ie\hbar A_0}{m_0 c}e^{-i(\mathbf{gr})}(\mathbf{a}^0\nabla)e^{i(\mathbf{k}_2\mathbf{r})}\varphi_{n'\mathbf{k}_2}\,d\tau =$$
$$= -\frac{eA_0}{m_0 c}\int e^{-i(\mathbf{k}_1\mathbf{r})}\varphi^*_{n\mathbf{k}_1}e^{-i(\mathbf{gr})}(\mathbf{a}^0\hat{\mathbf{p}})e^{i(\mathbf{k}_2\mathbf{r})}\varphi_{n'\mathbf{k}_2}\,d\tau. \qquad (76.44)$$

Taking into account that $\nabla\psi_{n'\mathbf{k}_2} = i\mathbf{k}_2\psi_{n'\mathbf{k}_2} + e^{i(\mathbf{k}_2\mathbf{r})}\nabla\varphi_{n'\mathbf{k}_2}$ we obtain

$$W_{n\mathbf{k}_1, n'\mathbf{k}_2} = \frac{i\hbar eA_0}{m_0 c}\int e^{-i(\mathbf{k}_1+\mathbf{g}-\mathbf{k}_2, \mathbf{r})}\varphi^*_{n\mathbf{k}_1}(\mathbf{a}^0, i\mathbf{k}_2)\varphi_{n'\mathbf{k}_2}\,d\tau +$$
$$+ \frac{i\hbar eA_0}{m_0 c}\int e^{-i(\mathbf{k}_1+\mathbf{g}-\mathbf{k}_2. \mathbf{r})}\varphi^*_{n\mathbf{k}_1}(\mathbf{a}^0\nabla)\varphi_{n'\mathbf{k}_2}\,d\tau. \qquad (76.45)$$

Since $|\mathbf{g}| \ll (\mathbf{k}_1, \mathbf{k}_2)$, the first term yields

$$\mathbf{k}_1 = \mathbf{k}_2 \text{ or } \frac{e\hbar(\mathbf{A}_0\mathbf{k}_2)}{m_1 c}, \qquad (76.46)$$

i.e. the result identical to the one we obtained with the aid of the de Broglie waves. The condition $\mathbf{k}_1 = \mathbf{k}_2$ reflects the quasi-momentum conservation law. Therefore we shall assume it to be valid for the second term of (76.45), as well. Combining both terms of (76.45) introduce the notation

$$P_{nn'}(\mathbf{k}_2)=\int \varphi^*_{n\mathbf{k}_2}(\mathbf{a}^0,\ -i\hbar\nabla+\hbar\mathbf{k}_2)\,\varphi_{n'\mathbf{k}_2}\,d\tau \qquad (76.47)$$

which will enable the matrix element of the perturbation operator to be written in the form

$$W_{n\mathbf{k}_1,\,n'\mathbf{k}_2}=W_{nn'}(\mathbf{k}_2)=-\frac{eA_0}{m_0c}P_{nn'}(\mathbf{k}_2) \qquad (76.48)$$

and the transition probability in the form

$$w(\mathbf{k}_1,\ \mathbf{k}_2)=\frac{2\pi}{\hbar}\frac{e^2A_0^2}{m_0^2c^2}\,|P_{nn'}(\mathbf{k}_2)|^2\,\delta_{\mathbf{k}_1\mathbf{k}_2}\delta[E(\mathbf{k}_1)-E(\mathbf{k}_2)+\hbar\omega]. \qquad (76.49)$$

Assuming the extrema to be located at the point $\mathbf{k}=0$ expand $P_{nn'}(\mathbf{k}_2)$ into a series

$$P_{nn'}(\mathbf{k}_2)=P_{nn'}(0)+\frac{dP_{nn'}(0)}{d\mathbf{k}_2}\mathbf{k}_2+\ldots \qquad (76.50)$$

In the effective mass approximation the first term equals zero, the second term, according to (76.17), will be equal to $(\mathbf{a}^0,\ \hbar\mathbf{k}_2)\frac{m_0}{m_p^*}$.

Making use of (76.30), (76.28), and (76.22) we obtain

$$\alpha=\frac{\delta q}{q}=\frac{1}{q}\int\int w(\mathbf{k}_1,\ \mathbf{k}_2)\,\frac{d\tau_{\mathbf{k}_1}}{4\pi^3}\,\frac{d\tau_{\mathbf{k}_2}}{4\pi^3}=$$
$$=\frac{8\pi\hbar nc}{\varepsilon\omega A_0^2}\cdot\frac{2\pi}{\hbar}\,\frac{e^2A_0^2}{c^2m_0^2}\int\int|P_{nn'}\,\mathbf{k}_2|^2\,\delta_{\mathbf{k}_1\mathbf{k}_2}\times$$
$$\times\delta[E(\mathbf{k}_1)-E(\mathbf{k}_2)+\hbar\omega]\frac{d\tau_{\mathbf{k}1}}{4\pi^3}\,\frac{d\tau_{\mathbf{k}2}}{4\pi^3}=\frac{e^2}{cm_0^2\pi^4n\omega}\times$$
$$\times\int|P_{nn'}(\mathbf{k}_2)|^2\,\delta[E_1(\mathbf{k}_2)-E_2(\mathbf{k}_2)+\hbar\omega]\,d\tau_{\mathbf{k}_2}. \qquad (76.51)$$

The transitions are termed allowed if $P_{nn'}(0)\neq0$. For such transitions, using the first term in (76.50) we obtain

$$\alpha=\frac{e^2}{\pi^4cm_0^2}\,\frac{|P_{nn'}(0)|^2}{n\omega}\int\delta[E_1(\mathbf{k}_2)-E_2(\mathbf{k}_2)+\hbar\omega]\,d\tau_{\mathbf{k}_2}. \qquad (76.52)$$

Using (76.33) and (76.36) calculate the integral of (76.52)

$$\int\delta[E_1(\mathbf{k}_2)-E_2(\mathbf{k}_2)+\hbar\omega]\,d\tau_{\mathbf{k}_2}=\int\limits_0^\infty\delta\left[(\hbar\omega-\Delta E_0)-\frac{\hbar^2\mathbf{k}_2^2}{2m^*_{red}}\right]4\pi k_2^2\,dk_2=$$
$$=2\pi\left(\frac{2m^*_{red}}{\hbar^2}\right)^{3/2}(\hbar\omega-\Delta E_0)^{1/2}, \qquad (76.53)$$

and obtain the expression for α:

$$\alpha = \frac{e^2}{\pi^4 c m_0^2} \frac{|P_{nn'}(0)|^2}{n\omega} \frac{2\pi (2m_{red}^*)^{3/2}}{\hbar^3} (\hbar\omega - \Delta E_0)^{1/2} =$$

$$= \frac{2e^2}{\pi^3 c m_0^2 \hbar^3} \frac{|P_{nn'}(0)|^2 (2m_{red}^*)^{3/2}}{n\omega} (\hbar\omega - \Delta E_0)^{1/2} =$$

$$= B (\hbar\omega - \Delta E_0)^{1/2}, \qquad (76.54)$$

where

$$B = \frac{2e^2}{\pi^3 c m_0^2 \hbar^3} \frac{|P_{nn'}(0)|^2 (2m_{red}^*)^{3/2}}{n\omega}. \qquad (76.55)$$

If $P_{nn'}(0) = 0$, the transitions are termed forbidden. Calculating α on the basis of (76.51) with the linear term of (76.50) taken into account we obtain an expression of the type of (76.39) but with a somewhat different value of A. Should the subsequent terms of the expansion (76.50) be taken into account, the expression $\alpha \sim (\hbar\omega - \Delta E_0)^{m+1/2}$ would be obtained, where m is the order of the derivative of the matrix element with respect to the wave vector. Generally, various "multiple" transitions may take place. However, their contribution will not be the same. To check this assess the ratio of the absorption coefficients corresponding to the allowed (76.54) and to the forbidden transitions (76.39):

$$\frac{\alpha_{for}}{\alpha_{al}} = \frac{A(\hbar\omega - \Delta E_0)}{B} = \frac{(2m_{red}^*)\, m_0^2 (\hbar\omega - \Delta E_0)}{3m_p^{*2} |(P_{nn'}(0)|^2}. \qquad (76.56)$$

It follows that for any finite value of $P_{nn'}(0)$ the pattern of the intrinsic absorption limit is determined by the allowed transitions. However, as the energy $(\hbar\omega - \Delta E_0)$ increases so does the contribution of the forbidden transitions. Thus, the dependence of α on $(\hbar\omega - \Delta E_0)$ for direct transitions should be of the form $\alpha = \alpha_0 (\hbar\omega - \Delta E_0)^r$, where r assumes the value 1/2 for allowed and 3/2 for forbidden direct transitions.

The absorption limit ω_{lim} determines the optical forbidden band width $\Delta E_0^O = \hbar\omega_{lim}$. The value of ΔE_0^O exceeds the minimum distance between the valence and the conduction bands which controls the thermal excitation of electrons.

The existence of states separated by a smaller energy gap than ΔE_0^O poses the question as to the possibility of the absorption of photons with lesser energy than $\hbar\omega_{lim}$. Obviously, in this case the selection rules (76.13-14) would be violated. However, the violation of the selection rule $\mathbf{k}_2 = \mathbf{k}_1 + \mathbf{g}$ is not tantamount to the violation of the quasimomentum (momentum) conservation law. The transition of an electron from the state $\mathbf{k}_1 \simeq 0$ to the state

$k_2 \cong k_0$ will be possible if the change in the electron momentum is compensated by the change in the momentum of the photon.

Consider the following processes. Let the electron be in the state $k_1=0$. Move it to the state $k_2=0$ by imparting to it the energy ΔE_0^O. Such an electron will become "hot" and will surrender its energy to the lattice as a result of collisions with it, creating photons with the energy $E_2(0)-E_c$ and the quasimomentum $\hbar K = -\hbar k_0$. The electron itself will go over to the state $E_2(k_0)$, its energy being altered by the amount $E_2(k_0)-E_1(0)=\Delta E_0^T < \Delta E_0^O$. Such a transition may be modelled as follows: the electron goes over from the state $E_1(0)$ to the state $E_2(k_0)$ following the absorption of a photon with the energy ΔE_0^T and of a long-wave phonon with the energy $[\Delta E_0^O - \Delta E_0^T]$ (this puts the electron in the state $E_2(0)$) and the emission of a phonon with the energy $\Delta E_0^O - \Delta E_0^T$ and the wave vector k_0 (this puts the eleótron in the state $E_2(k_0)$). In this way the electron goes over from $E_1(0)$ to $E_2(k_0)$ absorbing a photon with the energy $\hbar\omega = \Delta E_0^T$. The energy $\Delta E_0^O - \Delta E_0^T$ needed to transfer the electron is provided by the lattice and is given back to it. The electron transition takes place via an intermediate state in which the long-wave phonon is transformed into a short-wave phonon. In other words, *the transition of the electron from the conduction band to the valence band takes place at the expense of the photon energy, the change of the electron momentum being compensated by the lattice (by a phonon).*

The model discussed above is not the only one possible. Indeed, an electron in the state $E_1(0)$ may absorb a phonon with the energy $(\Delta E_0^O - \Delta E_0^T)$ and the quasimomentum $\hbar k_0$ and go over to some virtual state to be followed by the absorption of a long-wave photon with the energy $\Delta E_0^O - \Delta E_0^T$ and a transition to the $E_2(k_0)$ state. The electron in the $E_1(0)$ state may emit a phonon with the energy $\Delta E_0^O - \Delta E_0^T$ and the momentum $(-\hbar k_0)$ and go over to some virtual state to be followed by the absorption of a photon with the energy ΔE_0^T and a transition to the state $E_2(k_0)$. Thus, the transition of the electron from the state $E_1(k_1)$ to the state $E_2(k_2)$ for $k_1 \cong 0$ and $k_2 \cong k_0$ proceeds via several virtual states. To arrive at the frequency dependence of α one should take into account the energy and quasimomentum conservation laws:

$$E_2(k_2) = E_1(k_1) + \hbar\omega \pm \hbar\omega_{phon}, \qquad (76.57)$$

$$k_2 = k_1 \pm K_{phon}, \qquad (76.58)$$

where ω_{phon}, K_{phon} are frequency and the wave vector of the absorbed (plus) or the radiated (minus) phonon. The condition for the absorption boundary should be

$$\hbar\omega_{lim} = E_2(k_0) - E_1(0) \pm \hbar\omega_{vhon} = \Delta E_0^T \pm \hbar\omega_{phon}. \qquad (76.59)$$

Thus, there are two boundaries of intrinsic absorption, the minimum optical forbidden band width for indirect non-vertical transitions being less than the thermal forbidden band width ΔE_0^T by the energy of the phonon $\hbar\omega_{phon}$.

If we presume now that the perturbation includes some characteristics of the phonons, we shall have to determine the electron transition probability both by the matrix element of the perturbation due to the electromagnetic field and by the matrix element of the perturbation due to the lattice.

Assuming the matrix element of the perturbation due to the lattice to be independent of the phonon frequency we obtain the frequency dependence of the absorption coefficient in the form

$$\alpha(\omega) \sim (\hbar\omega - \Delta E_0^T \pm \hbar\omega_{phon})^2. \qquad (76.60)$$

Since the number of phonons depends on their energy and temperature, the expression for α may be written in the form

$$\alpha(\omega) \sim \left\{ \frac{(\hbar\omega - \Delta E_0^T + \hbar\omega_{phon})^2}{e^{\frac{\hbar\omega_{phon}}{kT}} - 1} + \frac{(\hbar\omega - \Delta E_0^T - \hbar\omega_{phon})^2}{1 - e^{-\frac{\hbar\omega_{phon}}{kT}}} \right\} \qquad (76.61)$$

The first term describes the light absorption process accompanied by the absorption of phonons whose number is proportional to $\frac{1}{e^{\frac{\hbar\omega_{phon}}{kT}} - 1}$, the second term describes the photon absorption process accompanied by the emission of a phonon, the probability of phonon emission being proportional to the probability that the non-excited oscillatory state in question is

$$1 - \frac{1}{e^{\frac{\hbar\omega_{phon}}{kT}} - 1} = \frac{1}{1 - e^{-\frac{\hbar\omega_{phon}}{kT}}}. \qquad (76.62)$$

For forbidden indirect transitions the power should be by a unit greater as compared to the power for indirect allowed transitions, i.e. instead of (76.60) we should write

$$\alpha(\omega) \sim (\hbar\omega - \Delta E_0^T \pm \hbar\omega_{phon})^3. \qquad (76.63)$$

Now let us compare the absorption coefficients resulting from direct and indirect transitions. The direct transition is conditional upon the probability of two particles meeting an electron and a photon. An indirect transition involves the meeting of three particles: an electron, a photon, and a phonon. This means that the indirect transition is a less probable process than the direct. *There-*

fore the light absorption coefficient should be greater for the direct transitions than for the indirect ones.

Figure 116 shows a diagram of direct (*a*) and indirect (*b*) transitions. The points *1* and *2* represent virtual states. Electron transitions accompanying light absorption in InSb may serve as an example of direct transitions. Indirect transitions take place, for instance, in germanium and silicon.

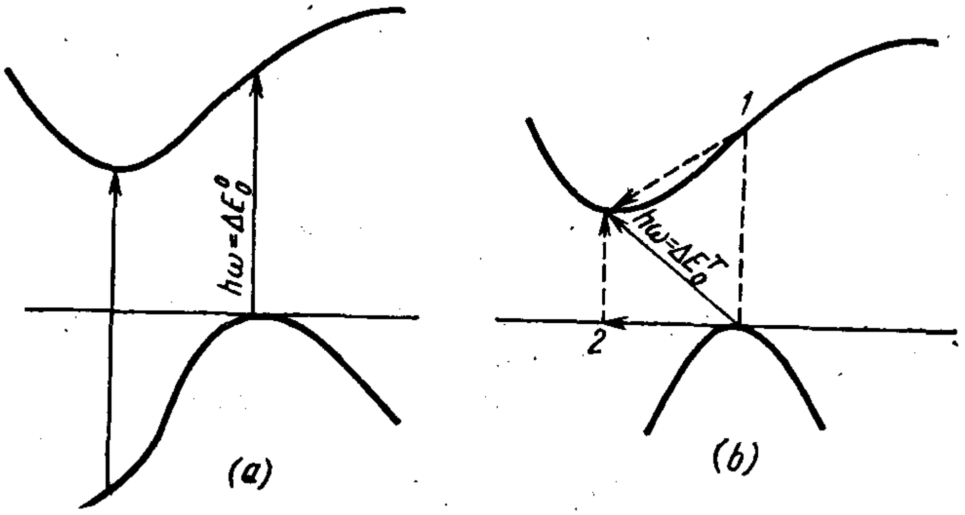


Fig. 116. Direct (*a*) and indirect (*b*) transitions in the course of intrinsic light absorption

Figure 117 shows absorption spectra for germanium and silicon. The intrinsic absorption boundaries which determine the optical forbidden band width are at 0.66 and 1.09 eV, respectively. The absorption edge is due to indirect transitions. In germanium there is a sharp increase in absorption in the region of 0.8 eV which may be explained by the direct transitions beginning to take place. The probability for such transitions is greater than for the indirect transitions. Direct transitions in silicon are observed for $\hbar\omega \geqslant 2.5$ eV.

Figure 118 shows in graphical form the dependence of intrinsic absorption in $A^{III}B^{V}$ on the wavelength. Figure 119 shows the same for some $A^{IV}B^{VI}$ compounds. A sharp increase in the absorption coefficient near the edge of the main band is due to the rapid increase in the number of particles capable of absorbing the photons as the energy of the latter rises. In Fig. 119 it may be seen that absorption beyond its intrinsic boundary is the absorption by charge carriers.

The study of the form of the absorption curve near the intrinsic absorption boundary in some cases enables a conclusion about the band pattern to be drawn.

Intrinsic absorption results in a rapid attenuation of light intensity since the coefficient of intrinsic absorption is of the order of 10^5-10^6 cm^{-1}.

In devising the theory of light absorption resulting in interband transitions of electrons we failed to take account of Coulomb interaction between holes and electrons created in the process of photon absorption. Coulomb attraction is instrumental in creating a bound electron-hole system, the exciton, which has a hydrogen-like discrete energy level system $E_N^{ex}=E_c-\frac{E_1^{ex}}{N^2}$ below the bottom of the conduction band (see Sec. 27). The fundamental state of the exciton, according to (27.9), is below E_c by the amount

$$E_1^{ex}=\frac{13.5}{\varepsilon^2}\left(\frac{m_{red}^*}{m_0}\right)(\text{eV}). \qquad (76.64)$$

In the course of direct interband transitions from the state $\mathbf{k}=\mathbf{k}_n$ a hole with the wave vector $\mathbf{k}_p=-\mathbf{k}$ is created. Since the exciton travels as a whole, it follows that the motion of the electron and the hole is correlated and that their relative velocity is zero. This is possible if the exciton springs from the transitions $\mathbf{k}_n=\mathbf{k}_p=0$, i.e. if the transitions take place in the centre of the Brillouin zone or, generally, in the energy extrema. The range of states from which allowed electron transitions resulting in the generation of excitons are possible is quite narrow, and this is the cause of formation of narrow absorption spectral bands adjoining the fundamental band from the long-wave side. Figure 120 shows an example of exciton absorption band observation in cadmium telluride and selenide. One band corresponding to the transition to the fundamental exciton state is observed in cadmium telluride, while three bands are observed in cadmium selenide corresponding to transitions from the three valence bands one of which is sunk by an amount of the order of 0.4 eV owing to the spin-orbital interaction (band C), and the other two originate from the atomic levels whose degeneracy has been removed by the action of the cadmium selenide wurtzite crystal lattice. The double degeneracy of the levels in cadmium telluride is not removed since its lattice is of the zinc blende type. The magnitude of the spin-orbital interaction in cadmium telluride is about 0.9 eV. The resolution of the band C, it being in the region of strong interband absorption, is poor.

The formation of narrow discrete absorption bands is not the only modification of the intrinsic absorption spectrum due to the exciton states. Calculations show the exciton states, i.e. Coulomb electron-hole interaction, to be responsible for the modification of the intrinsic absorption band pattern.

As is well known, every atomic system has an infinite set of discrete energy levels corresponding to a finite motion of the electron. When the potential energy of the interaction is normalized to be zero at infinity, the total energy of the electron is negative. For positive values of energy the electron is not bound to the ion and is moving freely, and the energy spectrum of its free motion is continuous. Overlapping transitions into discrete and continuous

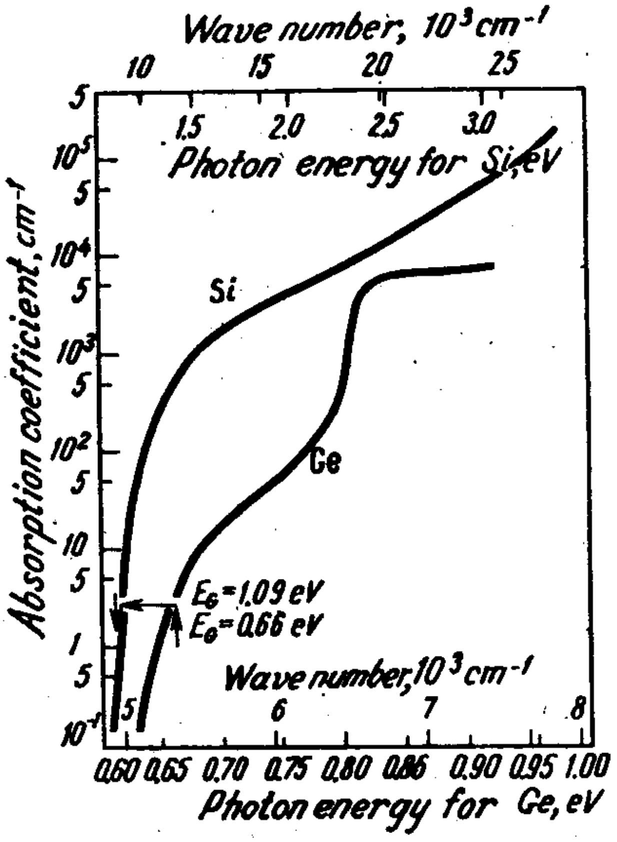


Fig 117. Absorption spectra of germanium and silicon

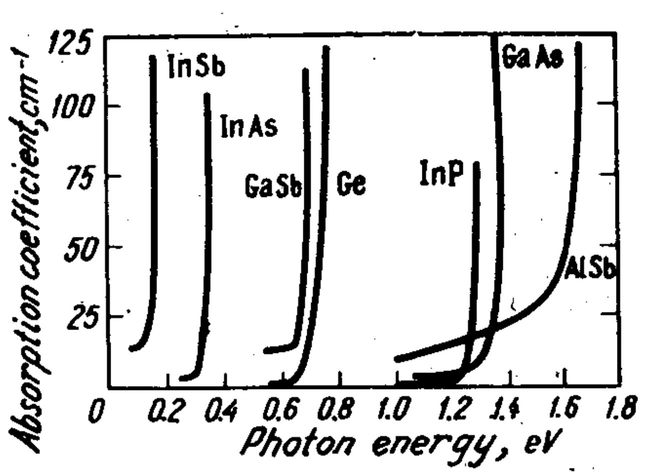


Fig. 118. Intrinsic absorption edge in $A^{III}B^{V}$ compounds and in germanium

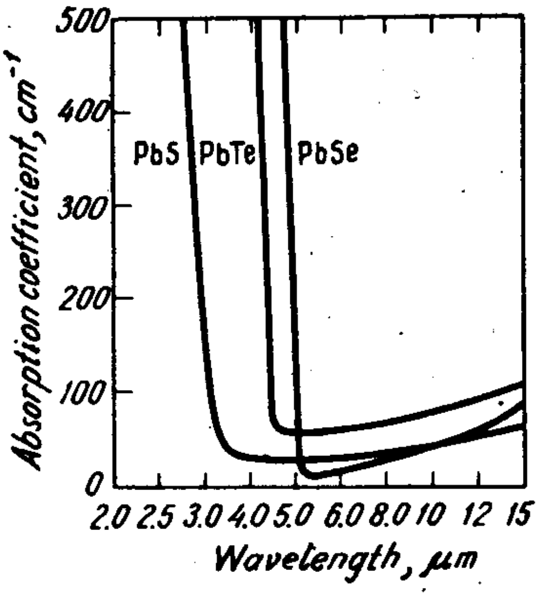


Fig. 119. Intrinsic absorption edge and free carrier absorption in $A^{IV}B^{VI}$ compounds

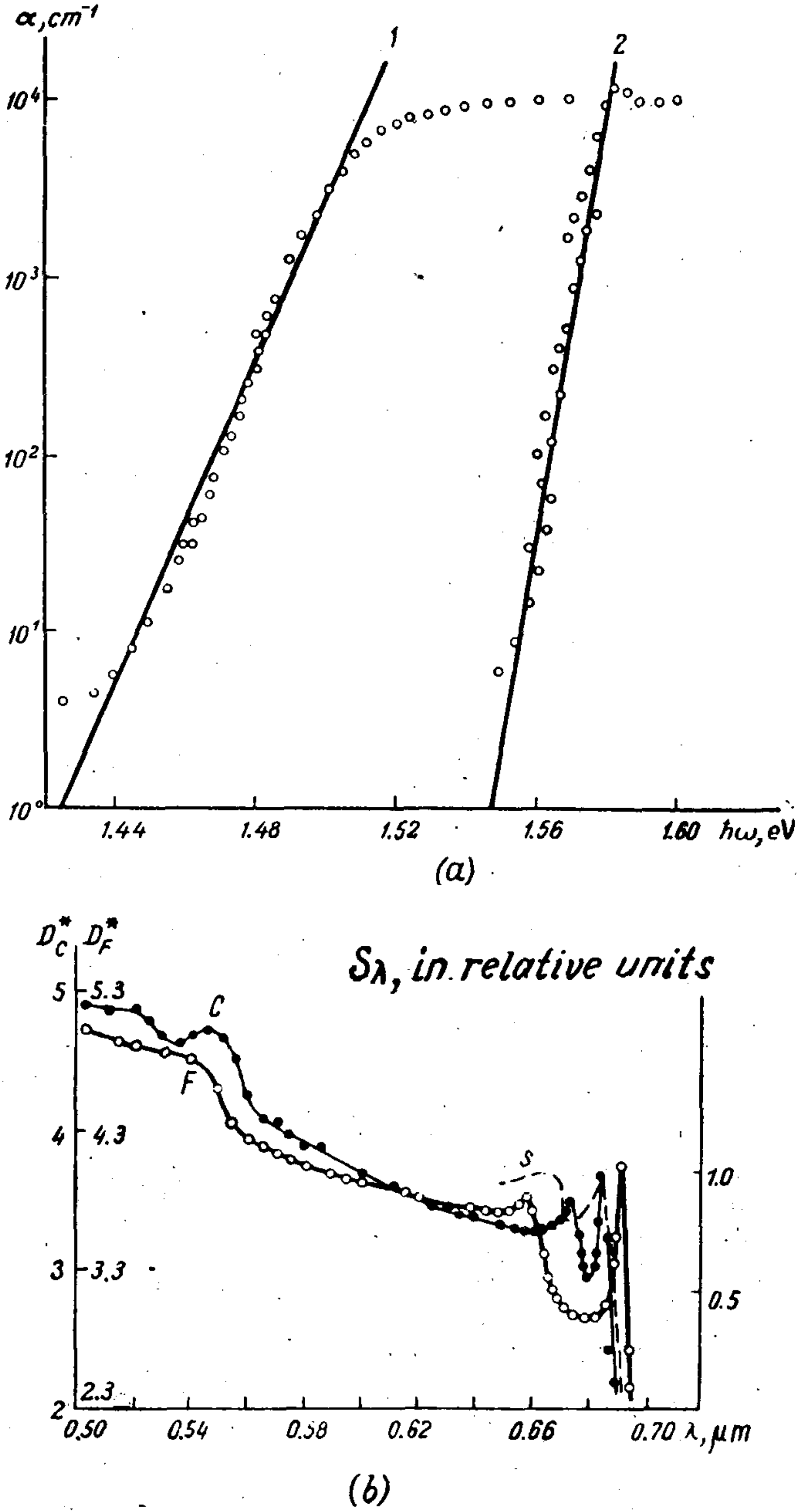


Fig. 120. Intrinsic absorption edge in cadmium telluride (*a*) and selenide (*b*). Absorption spectra of cadmium films on sapphire (curve *C*) and fluorite (curve *F*) substrates for 80 K and photosensitivity spectrum of a thick film (~20 µm) on a fluoride substrate (curve *S*)

regions of the energy spectrum prevent the absorption coefficient from turning zero when $\hbar\omega = \Delta E_0$. The calculations of Elliott, Dexter, McLean *et al.* result in the following absorption spectrum for direct allowed interband transitions:

$$\alpha = \frac{2\pi e^2}{m_0^2 c\omega n}\left(\frac{2m_{red}^*}{\hbar^2}\right)^{3/2} |P_{nn'}(0)|^2 (E_1^{ex})^{1/2} \frac{e^z}{\sinh z}, \qquad (76.65)$$

where

$$z = \pi\sqrt{\frac{E_1^{ex}}{\hbar\omega - \Delta E_0}}. \qquad (76.66)$$

For $\hbar\omega \longrightarrow \Delta E_0$ we obtain

$$\alpha(\Delta E_0) = \frac{4\pi e^2}{m_0^2 c\omega n}\left(\frac{2m_{red}^*}{\hbar^2}\right)^{3/2} |P_{nn'}(0)|^2 (E_1^{ex})^{1/2}, \qquad (76.67)$$

i.e. the greater the exciton ionization energy E_1^{ex} the greater is α for $\hbar\omega = \Delta E_0$. For $E_1^{ex} \longrightarrow 0$ $\alpha(\Delta E_0) \longrightarrow 0$ too, and the dependence $\alpha(\omega)$ which follows from the equation (76.65) takes the form $\alpha \sim \sim (\hbar\omega - \Delta E_0)^{1/2}$. For direct forbidden transitions the corrections obtained are similar. The theory of indirect exciton transitions has also been developed.

It follows from the exciton transition model that there should be a well-defined long-wave boundary of the fundamental band. Actually the long-wave edge of the fundamental absorption band is more or less spread out in the direction of $\hbar\omega < \Delta E_0$.

Figure 121 shows the dependence of the $\alpha^{1/2}(\hbar\omega)$ type for cadmium telluride based on the data of Fig. 120*a*. Several straight-line sections are clearly visible on the $\alpha^{1/2}(\hbar\omega)$ curve. Extrapolating these to $\alpha^{1/2} = 0$ we obtain energies separated by intervals of 0.021 eV which is close to the energy of the longitudinal optical phonon. This means that phonons interacting with excitons take part in the formation of the absorption edge. For indirect transitions only two straight-line sections should be observed on the $\alpha^{1/2}(\hbar\omega)$ curve which when extrapolated to $\alpha^{1/2} = 0$ should be separated by an energy interval equal to the double phonon energy. The exciton-phonon absorption theory is due to Segall.

The pattern of the intrinsic absorption band edge for the exciton-phonon interband transitions in many solids may be described by the Urbach equation valid for a wide range of α:

$$\alpha(\hbar\omega) = \alpha_0 e^{-\frac{\sigma(\hbar\omega - E_0)}{kT}} \qquad (76.68)$$

The parameter E_0 of the Urbach equation may be correlated with the energy maximum of the exciton absorption band the parameter α_0 coinciding with the value of the absorption coefficient in the maximum of the exciton band. The values of the parameter σ are in the range 1-3. Making use of the temperature dependence of σ one can describe the variation of the fundamental band-width temperature.

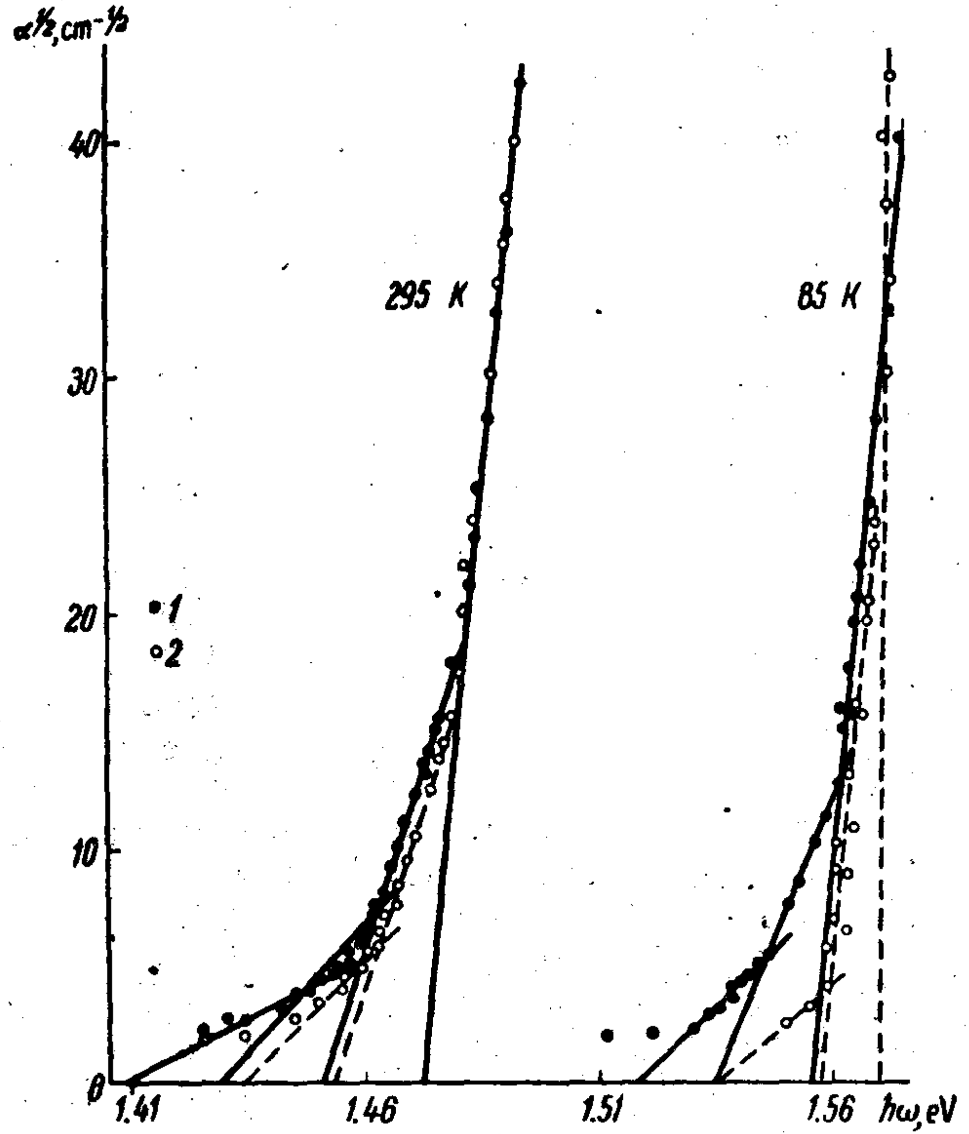


Fig. 121. The dependence of $\alpha^{1/2}$ on photon energy for cadmium telluride (calculated from data of Fig. 120*a*):

1—the original specimen; *2*—the specimen is annealed at 500° C

Summary of Sec. 76

1. Electron transitions in the process of light absorption are termed direct, or vertical, if the selection rules are satisfied:

$$\mathbf{k}_2 = \mathbf{k}_1 + \mathbf{g}, \quad \text{or} \quad \mathbf{k}_2 \cong \mathbf{k}_1. \tag{76.1s}$$

The transitions are termed indirect, or non-vertical, if

$$\mathbf{k}_2 \cong \mathbf{k}_1 + \mathbf{K}_{phon}. \tag{76.2s}$$

2. The relations determining the intrinsic band edge for direct and indirect transitions are

$$\hbar\omega_{lim} = \Delta E_0^O, \tag{76.3s}$$

$$\hbar\omega_{lim} = (\Delta E_0^T \pm \hbar\omega_{phon}). \tag{76.4s}$$

3. The coefficient α near the intrinsic absorption edge is proportional to the rth power of the difference $(\hbar\omega-\Delta E_0)$:

$$\alpha \sim (\hbar\omega-\Delta E_0)^r=(\hbar\omega-\hbar\omega_{lim})^r; \qquad (76.5s)$$

$r=1/2$ for direct allowed, 3/2 for direct forbidden, and 2—for indirect allowed, 3—for indirect forbidden transitions.

4. Intrinsic absorption results in the generation of an electron-hole pair.

The Coulomb interaction of the electron and hole generated by the light changes the pattern of the intrinsic edge for interband absorption with the result that the absorption coefficient at the band boundary as defined by the relations (76.3s) and (76.4s) does not turn zero.

5. Intrinsic absorption results in rapid attenuation of light, the mean free path of a photon l_{phon} for $\hbar\omega>\hbar\omega_{lim}$ being of the order of 10^{-5}-10^{-6} cm.

6. The perturbation operator is of the form

$$\hat{\mathbf{W}}=-\frac{1}{c}(\mathbf{A}\,\mathbf{j})=\frac{ie\hbar}{m^*c}(\mathbf{A}\nabla). \qquad (76.6s)$$

77. ABSORPTION OF LIGHT BY THE LATTICE

The absorption of light by the lattice is caused by the interaction of the light wave with the mobile charges in the lattice sites. In ionic crystals the localized ion charge with the coordinate $\mathbf{R}_{nj}$ and the velocity $\dot{\mathbf{R}}_{nj}$ interacts with the light wave which may be described by the vector potential $\mathbf{A}(\mathbf{r}, t)$. The atoms of homopolar crystals will have a dipole moment produced by the motion of the ions or induced by the light wave field. To be definite, we shall consider the motion of the lattice ions. The energy of interaction between light and the lattice according to (76.5) may be written in the form

$$\hat{\mathbf{W}}=\sum_{nj}\left\{-\frac{1}{c}e_j(\mathbf{A}[\mathbf{R}_{nj}]\dot{\mathbf{R}}_{nj})+\frac{e_j^2}{2M_jc^2}[\mathbf{A}(\mathbf{R}_{nj})]^2\right\}. \qquad (77.1)$$

For a wavelength greater than the lattice period, not to mention the displacement u_{nj}, one can take the value of the vector potential corresponding to $u_{nj}=0$, i.e. to the equilibrium position of the ions $\mathbf{R}^0_{nj}$. Assuming the light intensity to be not too large we may drop the second term of the expression (77.1). Introduce normal co-ordinates instead of $\dot{\mathbf{R}}_{nj}$:

$$\dot{\mathbf{R}}_{nj}=\dot{u}_{nj}=\sqrt{M_j}\,\dot{\mathbf{w}}_{nj}. \qquad (77.2)$$

The reduced displacements may be written in the form

$$\mathbf{w}_{\mathbf{n}j,\,\alpha(s)} = \mathbf{w}_{0j,\,\alpha(s)}\, e^{-i\,[\omega_{\alpha(s)}t-(\mathbf{K}_{\alpha(s)}\mathbf{n})]}. \qquad (77.3)$$

From (77.3) and (77.2) we write

$$\dot{\mathbf{R}}_{\mathbf{n}j} = -i \sum_{\alpha(s)} \mathbf{u}^{0}_{0j,\,\alpha(s)}\, e^{-i\,[\omega_{\alpha(s)}t-(\mathbf{K}_{\alpha(s)}\mathbf{n})]}\,\omega_{\alpha(s)}\,, \qquad (77.4)$$

therefore the perturbation operator (77.1) should in accordance with (77.4) be

$$\hat{\mathbf{W}} = -\frac{i}{c}\sum_{\mathbf{n}j}\sum_{\alpha(s)} e_j\, \omega_{\alpha(s)}\,(\mathbf{A}\,[\mathbf{n}+\mathbf{j},\, t]\,\mathbf{u}^{0}_{0j,\,\alpha(s)})\, e^{-i\,[\omega_{\alpha(s)}t-(\mathbf{K}_{\alpha(s)}\mathbf{n})]}. \qquad (77.5)$$

Write the vector potential

$$\mathbf{A}\,(\mathbf{n}+\mathbf{j},\, t) = \mathbf{A}_0 e^{i\,[\omega t-(\mathbf{g}\mathbf{R}^{0}_{\mathbf{n}j})]}. \qquad (77.6)$$

Before seeking the transition probability the states between which the transition takes place should be specified. We shall be interested in the phonon system characterized by the energy E and the wave function Ψ

$$E = \sum_{\alpha(s)} \hbar\omega_{\alpha(s)}\left(v_{\alpha(s)} + \frac{1}{2}\right). \qquad (77.7)$$

$$\Psi = \prod_{\alpha(s)} \Psi v_{\alpha(s)}\,(\mathbf{q}_{\alpha(s)}). \qquad (77.8)$$

To find the transition probability from one state (Ψ_1, E_1) to another, (Ψ_2, E_2), the matrix element of the perturbation operator should be found:

$$W_{12} = \int \Psi_1^{*}\,\hat{\mathbf{W}}\,\Psi_2\, d\Gamma', \qquad (77.9)$$

where $d\Gamma'$ is the volume element in the configurational space. Substituting the expression for $\hat{\mathbf{W}}$ (77.5) into (77.9) one may calculate W_{12}.

This method of calculation takes no account of the changes taking place in the photon system. To do so one should write the photon wave function using secondary quantization methods and introducing generation and annihilation operators. We shall confine ourselves to qualitative reasoning.

When calculating the matrix element W_{12} one should keep in mind that the integration should be carried out over all the normal co-ordinates; note, moreover, that $u_{0j,\,\alpha(s)}$ is a normal co ordinate so that $u_{0j,\,\alpha(s)} = \mathbf{q}_{\alpha(s)}$. Therefore, in the process of integration all the quantities except $u^0_{0j,\ \alpha(s)}$ may be taken out of the integral sign:

$$W_{12}=-\frac{i}{c}\sum_{nj}\sum_{\alpha(s)} e_j\omega_{\alpha(s)}\,\mathrm{A}_0\cos\theta_j e^{\pm i\,[\omega_{\alpha(s)}t-(\mathbf{K}_{\alpha(s)}\,\mathbf{n})]}\times$$
$$\times\int\psi^*v'_{\alpha(s)}(\mathrm{q}_{\alpha(s)})\,\mathrm{q}_{\alpha(s)}\,\psi v''_{\alpha(s)}(\mathrm{q}_{\alpha(s)})\,d\mathrm{q}_{\alpha(s)}. \qquad (77.10)$$

Other phonon functions when integrated yield unities.

The matrix element of the harmonic oscillator co-ordinate is equal to

$$\int\psi^*v'_{\alpha(s)}(\mathrm{q}_{\alpha(s)})\,\mathrm{q}_{\alpha(s)}\,\psi v''_{\alpha(s)}(\mathrm{q}_{\alpha(s)})\,d\mathrm{q}_{\alpha(s)}=$$
$$=\sqrt{\frac{\hbar}{2M_j\omega_{\alpha(s)}}}\begin{cases}\sqrt{v''_{\alpha(s)}}; & v'_{\alpha(s)}=v''_{\alpha(s)}-1;\\ \sqrt{v''_{\alpha(s)}+1}; & v'_{\alpha(s)}=v''_{\alpha(s)}+1.\end{cases} \qquad (77.11)$$

Hence, *the matrix element of the perturbation operator* W_{12} *is nonzero only when the variation of the quantum numbers is unity*: $v''=$ $=v'\pm 1$ *for all frequencies*. For $v''=v'-1$ the number of phonons *decreases* by unity: *the annihilation of a phonon* takes place; for $v''=v'+1$ *a phonon is generated* at the expense of a photon.

Substituting (77.11) into (77.10) we obtain

$$W_{12}=-\frac{i}{c}\sum_{nj}\sum_{\alpha(s)} e_j\sqrt{\frac{\hbar}{2M_j}}\,\omega^{1/2}_{\alpha(s)}\times$$
$$\times\sqrt{v'_{\alpha(s)}+1}\,\mathrm{A}_0\cos\theta_j e^{i\,(\omega-\omega_{\alpha(s)})\,t}\,e^{i\,(\mathbf{K}_{\alpha(s)}-\mathbf{g},\,\mathbf{n})}e^{-i\,(\mathbf{g}\mathbf{j})}. \qquad (77.12)$$

For the trànsition probability w (1, 2) we obtain

$$w(1,\ 2)=\frac{2\pi}{\hbar^2c^2}\mathrm{A}_0^2\Big|\sum_{j}\sum_{\alpha(s)} e_je^{-i\,(\mathbf{g}\mathbf{j})}\cos\theta_j\sqrt{\frac{\hbar\omega_{\alpha(s)}(v'_{\alpha(s)}+1)}{2M_j}}\times$$
$$\times\sum_{\mathbf{n}}e^{i\,(\mathbf{K}_{\alpha(s)}-\mathbf{g},\,\mathbf{n})}\Big|^2\times\delta(\omega-\omega_{\alpha(s)}). \qquad (77.13)$$

It follows that *the absorption of light by lattice vibrations is of a resonance nature*: $w(1,\ 2)=0$ *if* $\omega\neq\omega(s)$. From a continuous spectrum of radiation the lattice vibrations absorb only oscillations with resonance frequencies $\omega=\omega_{\alpha(s)}$. The values of $\omega_{\alpha(s)}$ are determined by all the optical and acoustical branches. Since $\omega_{\alpha(s)}=\omega$ the expression (77.13) may be substantially simplified. First, however, make use of the fact that the last sum in (77.13) is easily calculated:

$$\sum_{\mathbf{n}}e^{i\,(\mathbf{K}_{\alpha(s)}-\mathbf{g},\,\mathbf{n})}=N\delta_{\mathbf{K}'_{\alpha(s)},\,\mathbf{g}+2\pi\mathbf{b}}, \qquad (77.14)$$

here N is the number of cells in the crystal, and $\mathbf{b}$—the reciprocal lattice vector. The meaning of (77.14) is obvious: *in the process of light absorption the quasimomentum conservation law should be satis-*

fied:

$$\mathbf{K}_{\alpha(s)} = \mathbf{g} + 2\pi\mathbf{b} \tag{77.15}$$

Confining ourselves, as usual, to the main Brillouin zone for lattice vibrations ($\mathbf{b} = 0$) we obtain

$$\mathbf{K}_{\alpha(s)} = \mathbf{g}, \tag{77.16}$$

i.e. only such photons are absorbed whose momentum is equal to the momentum of the phonons. Combining the relations (77.13) and (77.15) we may write

$$\begin{aligned} \hbar\omega_{\alpha(s)} &= \hbar\omega, \\ \hbar\mathbf{K}_{\alpha(s)} &= \hbar\mathbf{g}, \end{aligned} \tag{77.17}$$

i.e. the energy and momentum conservation laws. They enable the spectrum of absorption by lattice vibrations to be found by means of geometrical ideas. Figure 122 shows two acoustical and two optical branches of lattice vibrations. The slope of the acoustical branches is equal to the speed of sound. Draw on the same diagram the light dispersion curve $\omega(\mathbf{g})$

$$\omega = cg. \tag{77.18}$$

The straight-line dependence of the photon frequency on the wave vector is shown in Fig. 122 by a dashed line. Since the slope should be equal to c (the velocity of light), the light dispersion curve should practically be a vertical line. Only the frequencies located at the intersection points of $\omega_{\alpha(s)}(\mathbf{K})$ and $\omega = cg$ will be absorbed because it is at these points that the energy and the momentum conservation laws are valid simultaneously.

It follows from here that *the acoustical branch does not actually absorb light* since it cannot intersect the straight line $\omega = cK$. *From each optical branch only the long-wave vibrations play a part in light absorption. In practice the spectrum of light absorption by the lattice vibrations may be assumed to consist of* $(3s - 3)$ $\omega^{0}_{(s)}(0)$ *frequencies.*

Now write the expression for the transition probability taking into account the selection rules (77.17) and (77.11):

$$w(1, 2) = \frac{\pi A_0^2 N^2 \hbar\omega}{\hbar^2 c^2} \left| \left[\sum_j e_j e^{-i(\mathbf{g},\, \mathbf{j})} \cos\theta_j \sqrt{\frac{v_0' + 1}{M_j}} \right] \right|^2. \tag{77.19}$$

Quantizing A_0^2 in the same way as it was done previously

$$A_0^2 = \frac{8\pi\hbar c n}{\varepsilon\omega}\, q, \tag{77.20}$$

we obtain

$$w(1,\,2)=\frac{8\pi^2 N^2 \mathrm{n}}{c\varepsilon}\,q\left|\sum_i e_j e^{-i\,(\mathbf{g}\cdot\mathbf{l})}\cos\theta_j\sqrt{\frac{v_0'+1}{M_j}}\right|^2. \qquad (77.21)$$

The effective cross section for the absorption of one photon is proportional to the number of phonons $v_0'+1$ of the frequency in question.

The order of magnitude of $\mathbf{l}$ is the same as that of the lattice constant, the wavelength in the absorption region being much greater, and by force of this $e^{-i\,(\mathbf{gl})}\cong 1$.

Consider now the dependence of $w(1,\,2)$ on the polarization of light and of the phonons. The angle between the polarization vectors of the photon and of the phonon is θ_j, and since $\mathbf{K}=\mathbf{g}$ for longitudinal optical vibrations, $\theta_j=90°$, and there is no absorption. *Only the transverse optical phonons will take part in absorption.*

When the crystal is illuminated by natural light the phonon and photon polarization may be decomposed into two orthogonal components, and this will enable us to use unity (with plus or minus sign) instead of $\cos\theta_j$. The sign of $\cos\theta_j$ should coincide with that of e_j (or be opposite to it). This is due to the fact that the displacements of the ions belonging to one cell are in antiphase. Taking note of that we write

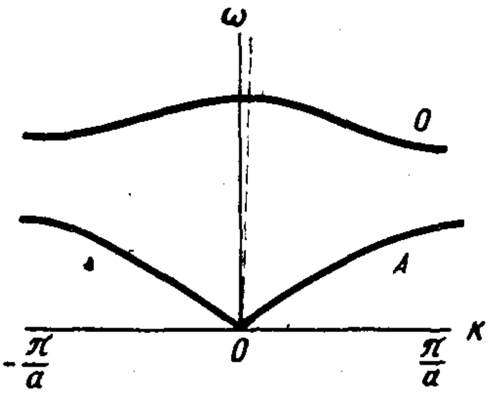


Fig. 122. Determination of the spectrum of light absorption by lattice vibrations

$$w(1,\,2)=\frac{8\pi^2 N^2 e^2 \mathrm{n}\,(v_0'+1)}{c\varepsilon M_{red}}\,q. \qquad (77.22)$$

We replaced M_j by some reduced mass M_{red}. For a two-ion lattice $M_{red}=\frac{M_1 M_2}{M_1+M_2}$. Putting $N(M_1+M_2)=\rho V$ we write

$$w(1,\,2)=\frac{8\pi^2 N V\rho e^2\,(v_0'+1)\,\mathrm{n}}{c\varepsilon M_1 M_2}\,q. \qquad (77.23)$$

Substitute for v_0' their equilibrium values

$$\langle v_0'\rangle=\frac{1}{e^{\frac{\hbar\omega}{kT}}-1}. \qquad (77.24)$$

The expression (77.24) establishes the frequency and the temperature dependence of the absorption of light by the lattice vibrations. As the temperature rises to $\frac{\hbar\omega}{kT}\gg 1$, $w(1,\,2)^{*}\sim T$. Assess the necessary temperature. Put the wavelength at 20 μm; in the case $\hbar\omega \cong 0.06$ eV which corresponds to the temperature of 400°C.

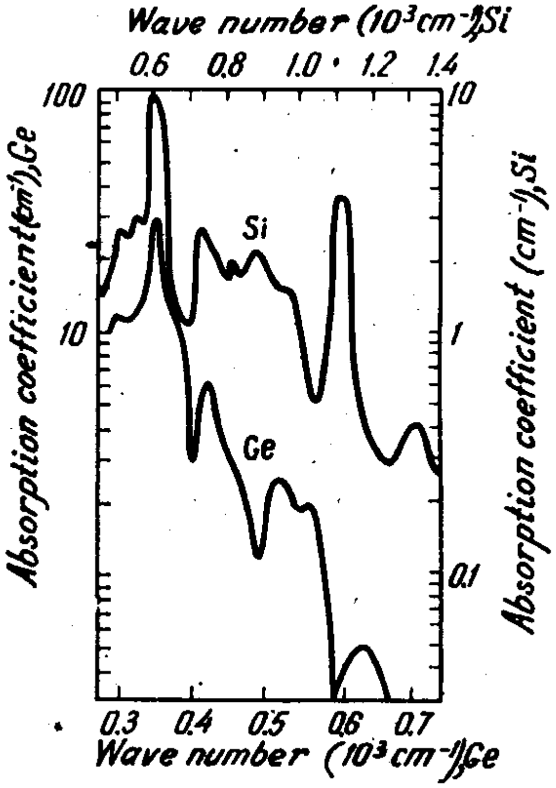


Fig. 123. Lattice vibration absorption spectrum in germanium and silicon

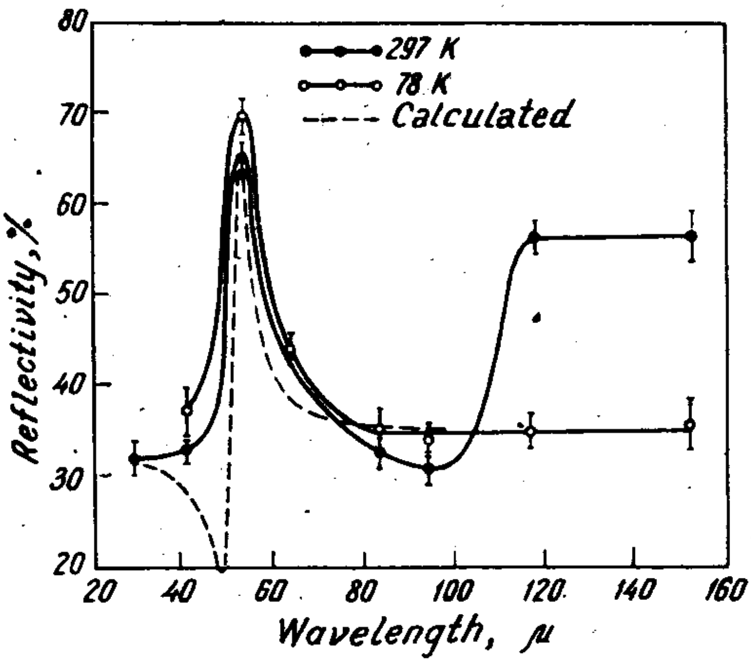


Fig. 124. Reflection spectrum of indium antimonide

Since the probabilities of direct and reverse transitions are equal, intense reflectivity should go in hand with intense absorption. The lattice absorption spectrum takes the form of the so-called *residual rays*. If the crystal is illuminated with a beam of a wide spectrum band, only the resonance frequencies of the lattice complying to the selection rules (77.17) will remain in the reflected beam. Actually the bands of the lattice vibration absorption spectrum are of a definite width. Figure 123 shows the lattice vibration absorption spectrum for germanium and silicon. Figure 124 shows the reflection spectrum of indium antimonide.

It may be seen from Fig. 123 that the lattice vibration absorption spectrum in germanium and silicon has several relatively wide bands in the region from 300 to 700 cm^{-1} (33-14 μm) in germanium and from 500 to 1400 cm^{-1} (20-7 μm) in silicon. The magnitude of the absorption coefficient measured in cm^{-1} is several units or tens. The absorption coefficient in germanium is larger than in silicon because its density is greater.

The reflection spectrum of indium antimonide contains a well-defined band in the region of $\lambda \cong 50$ μm which is due to the lattice vibrations. There are two points worthy of note in connection with Fig. 124. At low temperatures, when the carrier concentration is small, the nature of the reflectivity is purely dielectric, and it is determined by the refraqtion index. At room temperature there is a marked ~~reflection~~ due to the absorption by free charge carriers. The reflectivity rises with the wavelength as long as $\lambda < \lambda_b$; for $\lambda > \lambda_b$ it remains constant.

Summary of Sec. 77

1. Lattice vibrations result in light absorption in compliance with the selection rules:

$$\hbar\omega_{\alpha(s)} = \hbar\omega, \qquad (77.1s)$$

$$\hbar\mathbf{K} = \hbar\mathbf{g}. \qquad (77.2s)$$

2. It follows from the selection rules (77.1s-2s) that light is absorbed only by the optical lattice vibrations, so that

$$\omega \cong \omega^{O}_{\alpha(s)}(0). \qquad (77.3s)$$

Light is absorbed by the transverse optical lattice vibrations. The effect of lattice vibrations may be observed in the reflection spectrum in the form of "residual rays".

78. LIGHT ABSORPTION BY ELECTRONS IN LOCALIZED STATES

An electron or a hole in a localized state may absorb a photon, the absorption being followed by a transition to another localized or to a free state. Light absorption bands corresponding to carrier transitions from discrete levels lie beyond the intrinsic absorption region in the region of lower frequencies. The position of the absorption band may be determined from the relation:

$$\lambda_{loc}=\frac{1.24}{|E_{loc}-E_{extr}|}\left(\frac{\mu\text{m}}{\text{eV}}\right) \qquad (78.1)$$

Let us discuss the elementary theory of light absorption by an electron in a localized state. Denote the initial state by the index 1, and the final state by the index 2. Take the wave function of the localized state to be

$$\psi_1=Ce^{-\varkappa\rho}e^{-\frac{i}{\hbar}E_{loc}t}, \qquad (78.2)$$

where $\rho=|\rho|=|\mathbf{r}-\mathbf{R}_{loc}|$, E_{loc} being the energy of the localized level. The quantity $\frac{\varkappa^{-1}}{2}$ is equal to the distance at which the probability of locating the electron decreases e times. The function (78.2) is an analogue of the wave function of the s-state of a single-electron atomic system. The factor C may be found from the normalizing condition for the wave function:

$$1=|C|^2\int e^{-2\varkappa\rho}\,d\tau=|C|^2\,4\pi\int\limits_0^\infty e^{-2\varkappa\rho}\rho^2\,d\rho=\frac{\pi|C|^2}{\varkappa^3}, \qquad (78.3)$$

whence

$$C=C^*=\frac{\varkappa^{3/2}}{\sqrt{\pi}}. \qquad (78.4)$$

Take the wave function of the electron in the conduction band to be the wave function of the final state

$$\psi_2=\frac{1}{L^{3/2}}e^{\left[-\frac{iE_2t}{\hbar}+i(\mathbf{k}_2\mathbf{r})\right]}. \qquad (78.5)$$

Take the interaction energy in a conventional form

$$\hat{\mathbf{W}}=-\frac{ie\hbar}{cm^*}(\mathbf{A}\nabla)=-\frac{ie\hbar}{cm^*}e^{i[\omega t-(\mathbf{g},\mathbf{r})]}(\mathbf{A}_0\nabla). \qquad (78.6)$$

Find the matrix element W_{12}:

$$W_{12}=\int\psi_1^*\hat{\mathbf{W}}\psi_2\,d\tau=-\frac{ie\hbar}{cm^*}\frac{\varkappa^{3/2}}{\sqrt{\pi}L^{3/2}}e^{\frac{i}{\hbar}[E_{loc}+\hbar\omega-E_2]t}\times$$

$$\times\int e^{-\varkappa\rho}e^{-i(\mathbf{g},\mathbf{r})}(\mathbf{A}_0\nabla)e^{-i(\mathbf{k}_2\mathbf{r})}\,d\tau. \qquad (78.7)$$

Place the origin at the point $\mathbf{R}_{loc}$. In this case $|\rho|=|\mathbf{r}|$. Noting that

$$(\mathbf{A}_0\nabla)e^{i(\mathbf{k}_2\mathbf{r})}=i(\mathbf{A}_0\mathbf{k}_2)e^{i(\mathbf{k}_2\mathbf{r})}, \qquad (78.8)$$

we write

$$W_{12}=\frac{e\hbar\varkappa^{3/2}(\mathbf{A}_0\mathbf{k}_2)e^{\frac{i}{\hbar}[E_{loc}+\hbar\omega-E_2]t}}{cm^*\sqrt{\pi}L^{3/2}}\int e^{-\varkappa r+i(\mathbf{k}_2-\mathbf{g},\,\mathbf{r})}\,d\tau. \qquad (78.9)$$

Calculate the integral in (78.9) putting

$$\mathbf{k}_2-\mathbf{g}=\mathbf{Q}. \qquad (78.10)$$

Write

$$\int e^{-\varkappa r}e^{i(\mathbf{Qr})}d\tau=\int\limits_0^\infty e^{-\varkappa r}r^2\,dr\int\limits_0^\pi e^{iQr\cos\theta}\sin\theta\,d\theta\int\limits_0^{2\pi}d\varphi=$$

$$=\frac{2\pi}{iQ}\int\limits_0^\infty\left[e^{(-\varkappa+iQ)r}-e^{-(\varkappa+iQ)r}\right]r\,dr=$$

$$=\frac{2\pi}{iQ}\left[\frac{1}{(\varkappa-iQ)^2}-\frac{1}{(\varkappa+iQ)^2}\right]=\frac{8\pi\varkappa}{(\varkappa^2+Q^2)^2}. \qquad (78.11)$$

From (78.11) and (78.9) we obtain for the matrix element

$$W_{12}=\frac{8\sqrt{\pi}\,e\hbar\varkappa^{5/2}}{cm^*L^{3/2}}\,\frac{(\mathbf{A}_0\mathbf{k}_2)}{(\varkappa^2+Q^2)^2}e^{\frac{i}{\hbar}[E_{loc}+\hbar\omega-E_2]t} \qquad (78.12)$$

The probability of transition from state 1 to state 2 per unit time is

$$w(1,\,2)=\frac{128\pi^2\hbar e^2}{c^2m^{*2}L^3}\,\frac{\varkappa^5}{(\varkappa^2+Q^2)^4}\,A_0^2k_2^2\cos(\widehat{\mathbf{A}_0,\,\mathbf{k}_2})\,\delta[E_{loc}+\hbar\omega-E_2]. \qquad (78.13)$$

It follows from the energy conservation law

$$E_2=E_{loc}+\hbar\omega \qquad (78.14)$$

that light absorption by a discrete centre may take place for any energy $\hbar\omega\geqslant E_c-E_{loc}=E_l$. Substituting for A_0^2 its expression in terms of the photon flux (76.22) re-write

$$w(1,\,2)=\frac{2^{10}\pi^3\hbar^2e^2\mathrm{n}}{3cm^{*2}L^3\varepsilon\omega}\,\frac{\varkappa^5}{(\varkappa^2+Q^2)^4}k_2^2q=$$

$$=\frac{2^{11}\pi^3e^2\mathrm{n}}{3cm^*L^3\hbar\varepsilon}\,\frac{\varkappa^5}{(\varkappa^2+Q^2)^4}\,\frac{(\hbar\omega-E_l)}{\omega^2}J. \qquad (78.15)$$

Multiplying (78.15) by the concentration of localized absorption centres N_{loc} we obtain the absorption coefficient

$$\alpha_{loc}=\frac{2^{11}\pi^3e^2\mathrm{n}N_{loc}}{3cm^*\hbar L^3\varepsilon}\,\frac{\varkappa^5}{(\varkappa^2+Q^2)^4}\,\frac{(\hbar\omega-E_l)}{\omega^2}. \qquad (78.16)$$

Since according to (78.14) the energy of the electron may assume any value, no limitations are placed on its quasimomentum by the selection rules. To find the selection rules consider the quantity

$$\beta=\frac{\varkappa^5 k_2^2}{(\varkappa^2+Q^2)^4}=\frac{\varkappa^5 k_2^2}{[\varkappa^2+(\mathbf{k}_2-\mathbf{g})^2]^4}\,. \tag{78.17}$$

The quantity g as compared to k_2 is negligible because for light absorption by an electron on a discrete level $g \cong 10^3\ \mathrm{cm}^{-1}$, and the wave vector for conduction band electrons with an energy of the order of kT is as high as $10^7\ \mathrm{cm}^{-1}$. Therefore

$$\beta=\frac{\varkappa^5 k_2^2}{(\varkappa^2+k_2^2)^4}\,. \tag{78.18}$$

If we regard α as a function of k_2^2 it will be evident from (78.18) that α passes through a maximum the position of which may be found from the condition

$$\frac{d}{dk_2}\,\frac{\varkappa^5 k_2^2}{(\varkappa^2+k_2^2)^4}=(\varkappa^2+k_2^2)^4\,2k_2-4k_2^2(\varkappa^2+k_2^2)^3\,2k_2=0, \tag{78.19}$$

or

$$k_2=\frac{\varkappa}{\sqrt{3}}\,. \tag{78.20}$$

Because of a drastic fall in α for high k_2 ($\alpha \sim k_2^{-6}$ for $k_2>\varkappa$) we may assume the wave vector of the electron to vary inside the interval $0<k_2<\varkappa$. In other words, in the course of light absorption by electrons in localized states only such photons are absorbed that induce electron transitions to the state $E_2(\mathbf{k}_2)=$ $=E_{loc}+\hbar\omega$, where

$$k_2 \cong \varkappa. \tag{78.21}$$

The equation (78.21) is a sort of *a selection rule for light absorption by an electron in a localized state*. It may be easily demonstrated that this rule follows from *the uncertainty relation*. Indeed, the mean momentum of an electron, localized in a potential well the dimensions of which are of the order of $2\varkappa^{-1}$, is zero, but the uncertainty of the momentum is

$$\Delta p \cong \frac{\hbar}{2\varkappa^{-1}}=\frac{\hbar\varkappa}{2}\,. \tag{78.22}$$

When the electron goes over to a free state its momentum may be $\hbar k_2 \cong \Delta p$, i.e.

$$k_2 \cong \frac{\varkappa}{2} \tag{78.23}$$

in full agreement with (78.21). Thus, we arrive at the conclusion that *the absorption spectrum for light absorption by an electron in a localized state should be a comparatively wide but, nevertheless, finite band.* The band maximum is at the frequency determined by the condition

$$\hbar\omega = E_2(\mathbf{k}_2) - E_{loc} = E_I + \frac{\hbar^2\varkappa^2}{8m^*}. \qquad (78.24)$$

For an electron on an impurity level the photoionization energy, as determined from the impurity absorption maximum, exceeds the energy of thermal impurity ionization $E_I = E_c - E_{loc}$ by the amount of the order of $\frac{\hbar^2\varkappa^2}{8m^*}$.

The shape of the absorption coefficient curve may be assessed on the basis of the relations (78.16-15) if (78.14) is substituted for k_2^2:

$$\alpha \cong \frac{\hbar\omega - E_I}{\hbar^2\omega^2\,[\hbar^2\varkappa^2 + 2m^*(\hbar\omega - E_I)]^4}. \qquad (78.25)$$

As may be seen from (78.25), α increases linearly with the frequency for $\hbar\omega < 2E_I$, passes through a maximum, then decreases, at first slowly ($\sim\omega^{-1}$) but much faster in the end.

The theory of radiation (absorption) of the hydrogen atom may be applied to the localized states with a hydrogen-like spectrum. In this case the absorption coefficient may be written on the basis of the expression for the transition probability:

$$\alpha = N_{loc}\frac{2^{10}\pi^2\hbar e^2}{3ncm^*E_I}\left(\frac{E_I}{\hbar\omega}\right)^4 \frac{\exp\left(4\left(1-\frac{E_I}{\hbar\omega}\right)\arctan\left(1-\frac{E_I}{\hbar\omega}\right)\right]}{1-\exp\left[2\pi\left(\frac{E_I}{\hbar\omega}-1\right)\right]}, \qquad (78.26)$$

where n is the refraction index, E_I—the impurity ionization energy. For $m^* \cong 10^{-27}$ g n = 4, $E_I = 0.05$ eV

$$\sigma = \frac{\alpha}{N_{loc}} \cong 4\times10^{-16}\ \text{cm}^2.$$

Clearly, different calculation methods will result in somewhat different expressions for the absorption spectrum in the case of transitions between a band and a localized state. The difference is not only in method but reflects the difference in absorption mechanisms. The simplest way to demonstrate this is to consider the transition from the valence band to the discrete level of a localized state by changing places of ψ_1 and ψ_2 in the matrix element of the perturbation operator in (78.7).

The transitions described by the matrix element (78.9) are forbidden transitions. For the allowed transitions a matrix element of the perturbation operator independent of the wave vector should be introduced. Since its calculation with the aid of the method presented above is quite simple we shall not dwell on the subject.

The localized states may be of a different physical origin. The easiest way to produce them is by doping. In normal conditions the donor and acceptor impurities are completely ionized and for this reason cannot absorb light. However, at low temperatures, when the ionization is not complete, impurity absorption may be observed.

Figure 125 shows a wide photo-ionization band of boron atoms in silicon. The maximum of the absorption coefficient is at $\hbar\omega \cong \cong 0.055$ eV followed by a slow decrease to zero in the interval of $\delta\hbar\omega \cong 2E_i$. Several narrow absorption bands are superimposed on the linear region of the absorption coefficient. They may be explained by the existence of excited energy levels of the boron atoms lying inside the forbidden band as shown in Fig. 30, the narrow absorption bands corresponding to transitions of impurity atoms from the ground to the excited state. Figure 126 shows absorption bands of arsenic atoms in silicon corresponding to photo-ionization and to transition to the excited states.

Table 26 presents the ionization energies of impurity atoms for photo- and thermal ionization. It is noteworthy that the energy of photo-ionization is somewhat greater than that of thermal ionization. Should it be presumed that this difference is due to the uncertainty relation and to the selection rules which follow from it the value of $\varkappa$ obtained according to (78.24) for the difference between the photo- and the thermal ionization energy $\delta E_i \cong \cong 10^{-2}$ eV $= 1.6 \times 10^{-14}$ erg, and for $m^* = m$ would be $\varkappa \cong 10^7$ cm^{-1}.

Table 26

Impurity	Thermal ionization energy, eV	Optical ionization energy, eV
B	0.045	0.046
Al	0.057	0.067
Ga	0.065	0.071
In	0.160	0.154
P	0.044	0.0503
As	0.049	0.0533
Sb	0.039	0.0426

Figure 127 shows the absorption band of indium-doped p-germanium at 5 K.

Sometimes narrow absorption bands (lines) are observed beyond the long-wave intrinsic absorption edge. These lines may be due to deep or to shallow levels accepting electrons from the valence

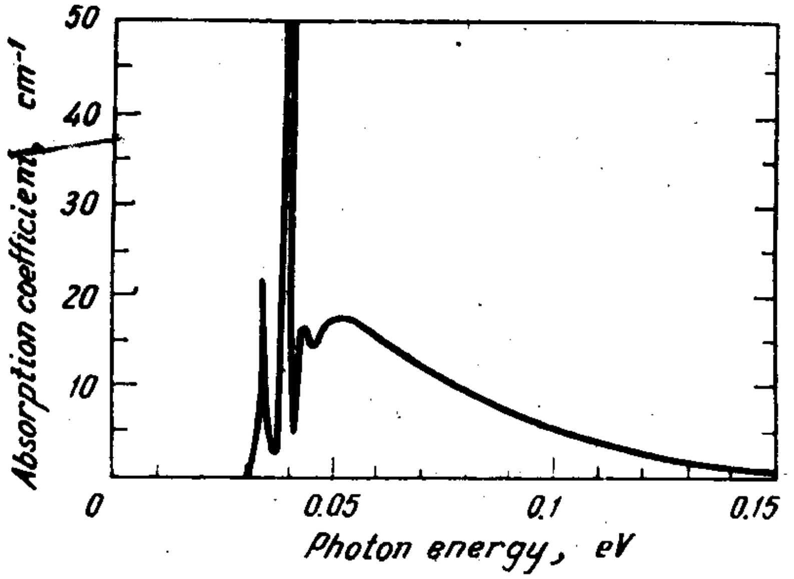


Fig 125. Absorption spectrum of boron atoms in silicon

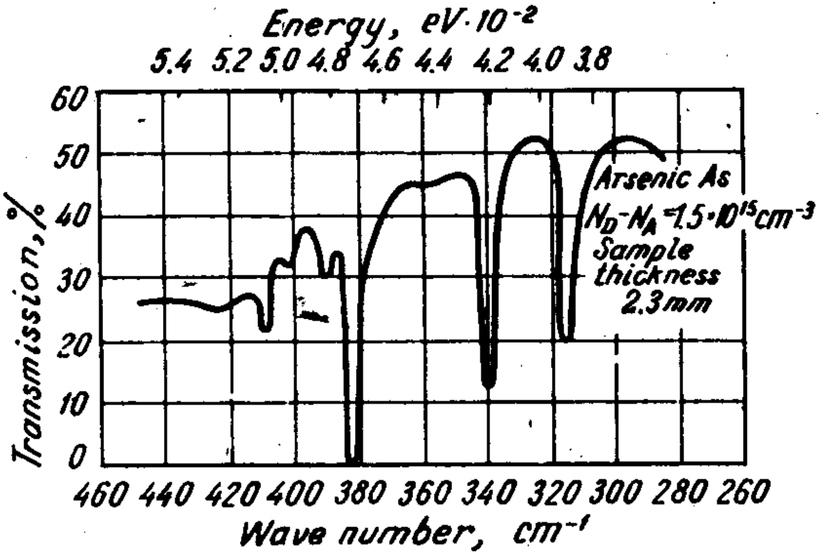


Fig. 126. impurity absorption spectrum of arsenic-doped silicon

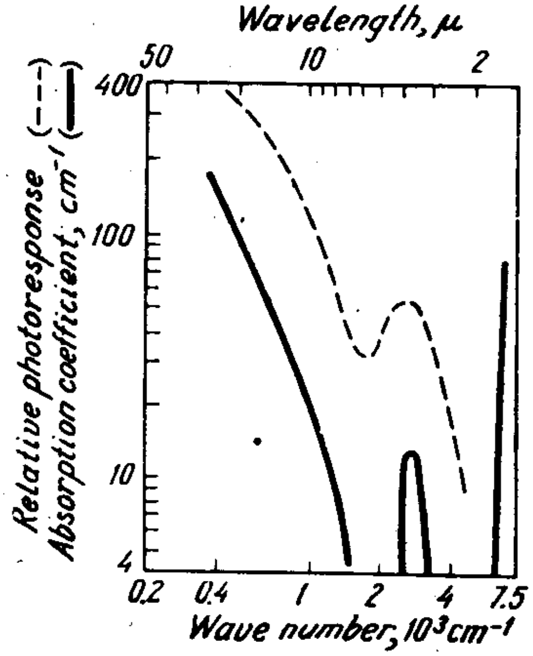


Fig. 127, Absorption coefficient of *p*-type germanium (solid line) and relative photoresponse of *p*-type indium-doped germanium at 5 K band. For some materials these lines may be interpreted as belonging to excitons. Exciton lines are observed in copper monoxide, cadmium sulphide, cadmium selenide, germanium, and in some other materials.

Figure 113 shows the absorption spectrum of neutron-irradiated *p*-silicon. The neutrons colliding with the silicon atoms produce a large number of defects. Irradiation results in substantial increase in resistivity of germanium and silicon. In Fig. 113 this is represented by a drastic reduction in the free carrier absorption. An absorption peak due to defects appears outside the boundaries of the intrinsic absorption band. The area bounded by the absorption curve is proportional to the irradiation dose which determines the defect concentration.

In ionic crystals transparent to visible light, electron-trapping defects serve as colour centres. An example of such defects is the *F*-centre—a vacancy that has trapped an electron.

In some cases the impurity is electrically neutral, i.e. it cannot be ionized thermally and, hence, cannot be a source of free carriers. But it can be ionized by light, this resulting in the appearance of absorption bands in the infrared region.

Oxygen in silicon may provide an example of such an impurity, for when bound in complexes it does not affect resistivity but produces an absorption band in the region of 9 μm.

Summary of Sec. 78

1. The selection rules for the absorption of light by charge carriers in localized states are of the form

$$\hbar\omega = E_2 - E_{loc} \tag{78.1s}$$

$$k_2 \cong \varkappa, \tag{78.2s}$$

where $2\varkappa^{-1}$ are the dimensions of the localization area of the electron or hole.

2. Impurity absorption takes the form of narrow bands if it is due to electron transitions between discrete energy levels or of comparatively wide bands if it is due to photo-ionization. The energy of photo-ionization is somewhat greater than the energy of thermal ionization.

3. Impurity absorption results in the generation of charge carriers of one type only.

79. INFLUENCE OF THE AMBIENT ON ABSORPTION SPECTRUM

The absorption spectrum of a semiconductor varies with the ambient conditions: temperature, pressure, and external fields.

The temperature affects the absorption spectrum in the following way. The rise in temperature is accompanied for most semiconductors by a reduction in the forbidden band width and in consequence the intrinsic absorption boundary shifts in the direction of longer waves. Figure 128 shows the absorption spectrum of selenium films at room (*1*, *2*, *3*) and at liquid nitrogen (*4*, *5*) temperatures. It is clearly visible that as the temperature changes from the liquid-nitrogen to room temperature the intrinsic absorption boundary shifts in the direction of longer waves.

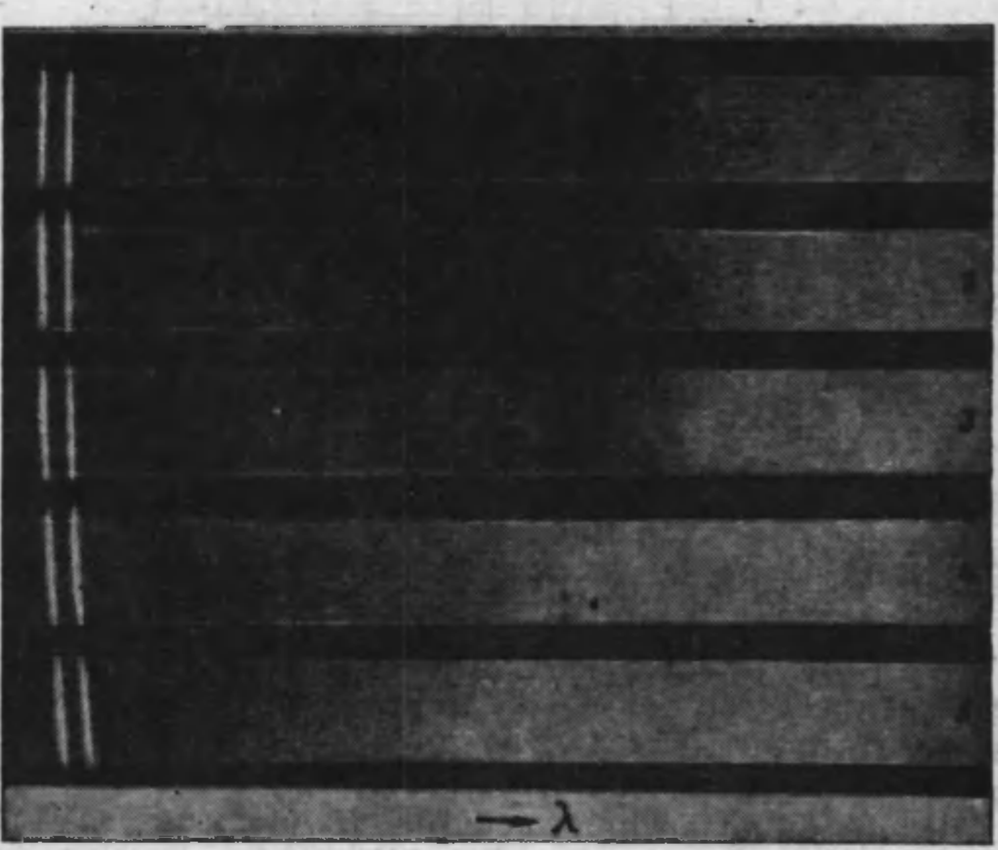


Fig. 128. The position of intrinsic absorption boundary in selenium films at room (*1-3*) and liquid nitrogen (*4-5*) temperatures

For an electrically active impurity the absorption spectrum may be observed only in the very low temperature range when the free charge carriers are "frozen" on the impurity level. In other words, as the impurity atoms become ionized with the rise in temperature the impurity absorption spectrum vanishes. Since the change in temperature is accompanied by a change in the free carrier concentration the absorption coefficient changes in the same direction, as may be seen from Fig. 114.

The temperature may indirectly affect the position of the intrinsic absorption boundary. If the density of states in the energy bands is small, they will rapidly be filled with carriers and unable to accommodate new carriers with the result that the intrinsic absorption boundary will shift to shorter waves. Usually, this is regarded as a concentration-dependent effect: the position of the intrinsic absorption boundary is determined by the impurity concentration. This is due to the fact that for high impurity concentration impurity ionization energy tends to zero. This effect is most noticeable in indium antimonide which has a small density of states in the conduction band. Figure 129 shows the dependence of the position of the intrinsic absorption boundary on electron concentration in InSb. As may be seen from Fig. 129, as the electron concentration changes from 10^{17} to 5×10^{18} cm^{-3}, the levels in the conduction band occupied by electrons shift to 0.3 eV above E_c The shift of the absorption edge accompanying the increase in the degree of degeneracy is known by the name of the Burstein effect.

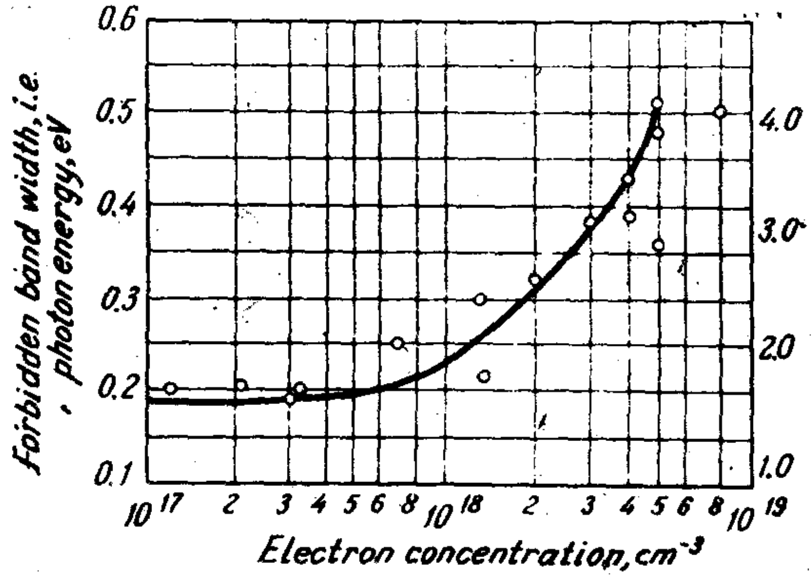


Fig. 129. The dependence of the position of intrinsic absorption boundary in InSb on electron concentration

Compression or extension of the crystal changes the forbidden band width and with it the position of the intrinsic absorption boundary.

This position is also affected by the electric field. Figure 130 shows the position of the intrinsic absorption boundary in a CdS film for three different voltages applied to the film. The intrinsic absorption boundary shifts to longer waves which is equivalent to a decrease in the forbidden band width in a strong electric field. This is *the Keldysh-Franz effect*.

Magnetic fields, too, affect the position of the intrinsic absorption boundary. It was shown in Chapters II and III that the application of a magnetic field raises the bottom of the conduction band by an amount $\frac{\hbar\omega_c}{2}=\frac{\hbar eB}{2m_n}$. The top of the valence band drops by the amount $\frac{\hbar eB}{2m_p^*}$. This leads to an increase in forbidden band width and to a corresponding shift of the intrinsic absorption boundary to shorter wavelengths:

$$\hbar\omega_{lim} = E_c(B) - E_v(B) = E_{c0} + \frac{\hbar eB}{2m_n^*} - \left(E_{v0} - \frac{\hbar eB}{2m_p^*}\right) =$$
$$= \Delta E_0 + \frac{\hbar eB}{2}\left(\frac{1}{m_n^*} + \frac{1}{m_p^*}\right) = \Delta E_0 + \frac{\hbar eB}{2m_{red}^*}. \qquad (79.1)$$

The measurement of the dependence of the intrinsic absorption boundary shift on the magnetic field may serve to determine the effective mass of the electron and hole. This, however, is not the only effect observed when light is absorbed in a semiconductor placed in a magnetic field. According to (76.36) the intrinsic absorption coefficient is proportional to the density of states in the energy bands. But, as was demonstrated in Sec. 35, this density changes with the application of a magnetic field—it becomes infinite on the Landau levels and practically zero in the intervals between the Landau parabolas (Fig. 58). This means, however, that the intrinsic absorption coefficient is not equal to zero only for the transitions between the Landau levels, and for this reason the absorption spectrum should be composed of a series of narrow bands. This effect observed in some materials at low temperatures and high magnetic fields is termed *oscillatory magneto-absorption effect*, or *magneto-absorption*. Figure 131*a* shows the ratio of the intensities of light transmitted by a germanium sample in the presence $J(B)$ and in the absence $J(0)$ of a magnetic field. We observe several peaks in the transmission spectrum whose position depends on the intensity of the magnetic field. Figure 131*b* shows the dependence of the position of absorption maxima on the magnetic field. For $B \to 0$ all the maxima contract into one point which corresponds to the optical forbidden band width for vertical transitions in germanium. This effect enables the optical forbidden band width to be determined with maximum accuracy. The oscillatory magneto-absorption effect is an analogue of the cyclotron resonance since in both cases the transitions are between the Landau levels. The difference is that in the cyclotron resonance electron transitions take place between the Landau levels of the same band, while *in the magneto-absorption effect the*

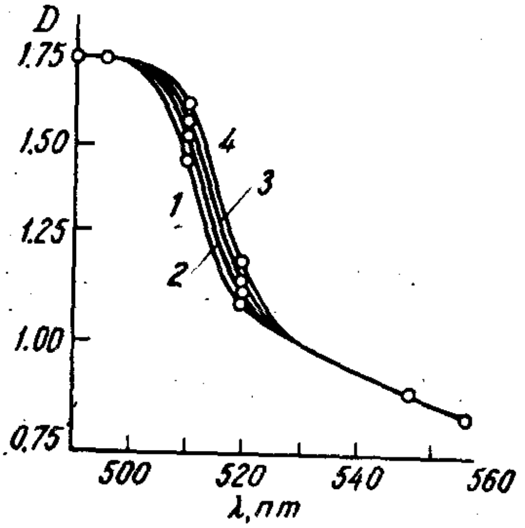


Fig. 130. The shift of intrinsic absorption edge in CdS film with the electric field voltage in kV: 1—0; 2—1.2; 3—1.6; 4—1.8

transitions are between the Landau levels of different bands. The difference in photon energies necessary for such transitions is of the order of the forbidden band width.

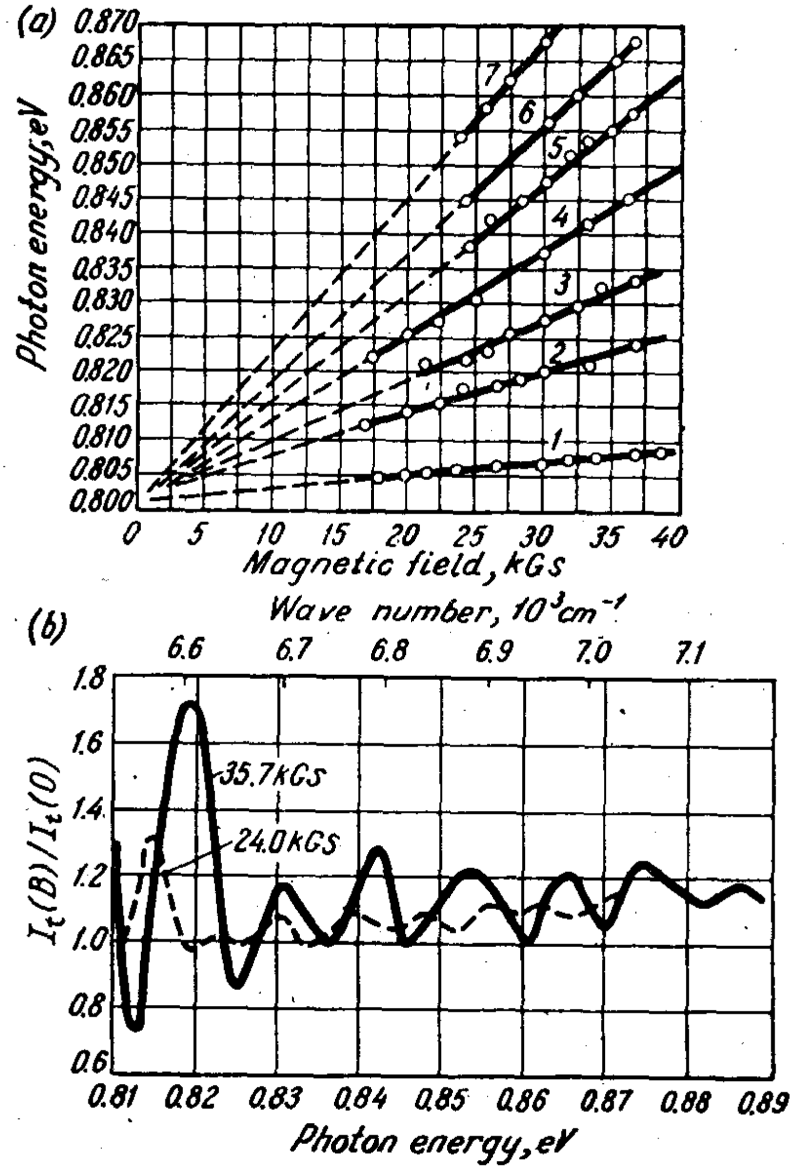


Fig. 131. Oscillatory magneto-absorption effect in germanium

80. PHOTORESISTIVE EFFECT

The variation of the electrical resistance of a semiconductor sample due to illumination is termed photoresistive effect, or internal photoeffect. The variation of the resistance or the conductance is caused by the variation of charge carrier concentration. The photoresistive effect may be described with the aid of photoconductivity σ_{pc}

$$\sigma_{pc} = e_n\mu_n\delta n + e_p\mu_p\delta p, \tag{80.1}$$

where δn, δp are excess electron and hole concentrations produced by illumination. If $\sigma_{pc} > 0$ the photoresistive effect is termed positive; if $\sigma_{pc} < 0$—negative. The conductivity $\sigma_0 = \sigma_d$ due to equilibrium carriers is termed dark-current conductivity. The total conductivity may be represented in the form of a sum of dark and light conductivities:

$$\sigma = e_n\mu_n (n_0 + \delta n) + e_p\mu_p (p + \delta p) = \sigma_d + \sigma_{pc}. \tag{80.2}$$

In the course of intrinsic absorption the electrons and holes are generated in equal numbers: $\delta n = \delta p$. In the course of exciton absorption a bound electron-hole pair is generated which is a neutron complex, therefore exciton absorption does not directly result in an increase of carrier concentration. However, if the exciton, having absorbed additional energy in the process of its motion in the crystal, dissociates this will result in the generation of two free carriers: an electron and a hole. The exciton concept itself, by the way, was introduced by Frenkel to account for the absence of the photoresistive effect in conditions of intense light absorption.

The photo-ionization of localized states such as impurity atoms or F-centres increases the concentration of charge carriers of one type only, there being two possible cases, the non-equilibrium carriers being of the majority or of the minority type. In case the non-equilibrium carriers are of the minority type and their concentration exceeds the dark concentration of the majority carriers, the result is the change of conductivity type. When the semiconductor is illuminated all kinetic phenomena may proceed in a way different from that in absence of illumination. For instance, the sign of the Hall coefficient may be reversed.

The increase in carrier concentration may be only a secondary result of the absorption of light by the lattice vibrations, the latter increasing the concentration of phonons capable of spending their energy on carrier excitation.

The absorption of light by free carriers does not change their concentration but causes deviations from the equilibrium distribution of carriers over the states, the carriers becoming "hot" with a corresponding change in their mobility. This, in turn, brings about changes in conductivity.

Up to now, when discussing the photoresistive effect, we had in mind an increase in conductivity. However, illumination of a semiconductor may also result in an increase in resistivity.

To describe the photoresistive effect one should know the light concentration of the charge carriers. This may be determined from the continuity equations (64.6) and (64.7):

$$\frac{\partial n}{\partial t} = -\operatorname{div}\frac{\mathbf{j}_n}{e_n} + G_n - \frac{n - n_0}{\tau_f^n}, \tag{80.3}$$

$$\frac{\partial p}{\partial t} = -\operatorname{div}\frac{\mathbf{j}_p}{e_p} + G_p - \frac{p - p_0}{\tau_f^p}. \tag{80.4}$$

Consider a semiconductor in the absence of current: $\mathbf{j}_n = \mathbf{j}_p = 0$. In this case the equations (80.3) and (80.4) may be simplified:

$$\frac{\partial n}{\partial t} = G_n - \frac{n - n_0}{\tau_f^n}, \tag{80.5}$$

$$\frac{\partial p}{\partial t} = G_p - \frac{p - p_0}{\tau_f^p}. \tag{80.6}$$

If the carrier generation rates G_n and G_p are known, the equations (80.5) and (80.6) may be used to find the carrier concentration and, consequently, σ_{pc}. In a stationary state $\frac{\partial n}{\partial t} = \frac{\partial p}{\partial t} = 0$, and the solution of the equations (80.5) and (80.6) is obvious:

$$n - n_0 = \delta n = G_n \tau_f^n, \tag{80.7}$$

$$p - p_0 = \delta p = G_p \tau_f^p. \tag{80.8}$$

The expressions (80.7) and (80.8) bear the name of *the first characteristic relation for the photoresistive effect*. They determine the dependence of photoconductivity on light intensity and on the spectral region. Write for the photoconductivity

$$\sigma_{pc} = e_p \mu_p (G_p \tau_f^p + b G_n \tau_f^n) = \Delta\sigma. \tag{80.9}$$

The value of σ_{pc} depends on the wavelength and the intensity of the incident light via the dependence of G and τ_f on λ and J. The generation rate G is directly affected by λ and J. The dependence of τ_f is via the dependence of the lifetime on excess carrier concentration, the latter being, in turn, dependent on the generation rate.

The generation rate G is determined by the light intensity J and the absorption coefficient α. Indeed, the amount of energy $-dJ$ absorbed in the volume 1 dx per unit time is

$$-dJ = \alpha J\, dx, \tag{80.10}$$

and per unit volume

$$-\frac{dJ}{dx} = \alpha J. \tag{80.11}$$

Expressing light intensity J in terms of the photon flux $q = \frac{J}{\hbar\omega}$ we obtain the number of photons absorbed in unit volume per unit time:

$$-\frac{1}{\hbar\omega}\frac{dJ}{dx} = (-)\, q_1 = \frac{\alpha J}{\hbar\omega} = \alpha q. \tag{80.12}$$

Thus, the *number of photons absorbed in unit volume per unit time q_1 is equal to the product of the coefficient α and the photon* *flux q*. Let each absorbed photon generate a free carrier (or a carrier pair) with a probability η. The quantity ηq_1 represents *the carrier generation rate:*

$$G_n = \eta_n q_1 = \eta_n \alpha q, \tag{80.13}$$

$$G_p = \eta_p q_1 = \eta_p \alpha q. \tag{80.14}$$

The quantities η_n and η_p are termed *quantum photo-ionization yields (efficiencies)*. In case of photoconductivity due to absorption by localized states one of the quantities η_n or η_p is zero. The value of η may be determined experimentally by measuring δn and δp. Its experimental values may be either greater or less than unity. This does not mean, however, thaf one photon can directly generate several free charge carriers. This only means that some secondary effects take place in the semiconductor with the result that the effective number of free charge carriers η_n and η_p per each quantum absorbed exceeds unity. Making use of the expression for the generation rate of charge carriers write the expression for current density due to photoconductivity, i.e. for the photocurrent density $\mathbf{j}_{pc}$:

$$\mathbf{j}_{pc} = \sigma_{pc}\mathbf{E} = e_p \mu_p (\eta_p \tau^p_f + b\eta_n \tau^n_f)\, \alpha q \mathbf{E}. \tag{80.15}$$

If the dimension of the semiconductor sample in the direction of the electric field is l, and the voltage across it is V, then $E = \frac{V}{l}$, $\mu_p E = v_{dp}$, $\mu_n E = v_{dn}$. The drift time $t_n = \frac{l}{v_{dn}}$ and $t_p = \frac{l}{v_{dp}}$. Express the field intensity in terms of the time in which the carriers traverse the sample:

$$E = \frac{v_{dp}}{\mu_p} = \frac{l}{\mu_p t_p} = -\frac{l}{\mu_n t_n}, \tag{80.16}$$

and substitute (80.17) into (80.16) to obtain

$$j_{pc} = \left(e_p \mu_p \eta_p \tau^p{}_f \frac{l}{\mu_p t_p} - e_n \mu_n \eta_n \tau^n_f \frac{l}{\mu_p t_n}\right) \alpha q =$$

$$= e_p \left(\eta_p \frac{\tau^p{}_f}{t_p} + \eta_n \frac{\tau^n_f}{t_n}\right) \alpha q l. \tag{80.17}$$

Multiply the expression (80.17) by the sample cross section S; Sl is the volume of the sample; $j_{pc} S = I$ is the photocurrent. For a sufficiently thin sample and for a sufficiently weak absorption αq will be the same at every point of the sample. Denote the total number of photons absorbed in the entire volume of the sample by q_v and introduce the notation

$$A' = \eta_n \frac{\tau^n_f}{t_n} + \eta_p \frac{\tau^p_f}{t_p}. \tag{80.18}$$

In this case the photocurrent is equal to

$$I_{pc}=e_p q_v A'. \tag{80.19}$$

If $\eta_n=\eta_p=\eta$ (or if one of the η's is zero),

$$q_v\eta=G \tag{80.20}$$

will be the total carrier generation rate, and

$$A'=\frac{\tau_f^n}{t_n}+\frac{\tau_f^p}{t_p} \tag{80.21}$$

is termed *amplification factor*. The expression for the photocurrent may be written in the form

$$I_{pc}=e_p GA'. \tag{80.22}$$

This is the second characteristic relation for the photoresistive effect.

For sufficiently high fields the transit time $t=\frac{l}{v_d}$ may turn out to be less than the lifetime. The expression (80.22) is valid only for ohmic contacts since only such contacts do not affect the composition of the current. The relation (80.21) shows how efficiently the photo-carriers are utilized.

If all the constituents of (80.17) except η are known, the quantum yield may be determined from measurements of I_{pc}. Figure 132*a* shows the spectral characteristic of quantum efficiency for germanium. This may be seen to be equal to unity up to 2.7 eV and to increase as the energy of the photon exceeds this value. This fact may be explained as follows. The quasimomenta of the free electron and the hole, generated by a direct transition of the valence band electron resulting from the absorption of a photon, are equal, their kinetic energies being inversely proportional to their effective masses. When the kinetic energy of one of the particles reaches the value of the forbidden band width, this "hot" particle is able to spend its energy on the generation of an additional free electron-hole pair. For $m_n^*\cong m_p^*$ the boundary of the quantum efficiency increase should lie in the region $\hbar\omega\cong 3\Delta E_0$. For $m_p^*\gg m_n^*$ this boundary shifts into the region $\hbar\omega\cong 2\Delta E_0$. As the temperature is raised the forbidden band width decreases, and the boundary of the increase in η shifts to lower energies as shown for silicon in Fig. 132*b*. Thus, the *increase of the quantum efficiency in excess of unity is a secondary effect caused by impact ionization and not by direct generation of two carrier pairs by one photon.*

For different recombination mechanisms there will be a different dependence of the photoconductivity on the light intensity. For linear recombination the excess carrier concentration is proportional to light intensity, and

$$\sigma_{pc} \sim J; \quad I_{pc} \sim J, \tag{80.23}$$

i.e. the photocurrent I_{pc} is proportional to the light intensity J. For quadratic recombination

$$\sigma_{pc} \sim \sqrt{J}, \quad I_{pc} \sim \sqrt{J}. \tag{80.24}$$

It may generally be assumed that

$$\sigma_{pc} \sim J^{\gamma}. \tag{80.25}$$

For $\gamma = 1$ the photoresistive effect is termed linear, for $\gamma < 1$—non-linear, and for $\gamma > 1$—superlinear.

However, even in the case of linear recombination the dependence of σ_{pc} or, δn_{st} on J is more intricate. As was demonstrated in Sec. 65 the lifetime τ_{∞} may be greater or less than τ_0, where τ_0 is the lifetime of a carrier pair for low, and τ_{∞}—for high generation levels. Therefore, $\delta n \sim J$ both in the low and high intensity ranges. But since the slope of the lines is different in both ranges, the dependence of δn on J in the intermediate intensity range may be non-linear or superlinear.

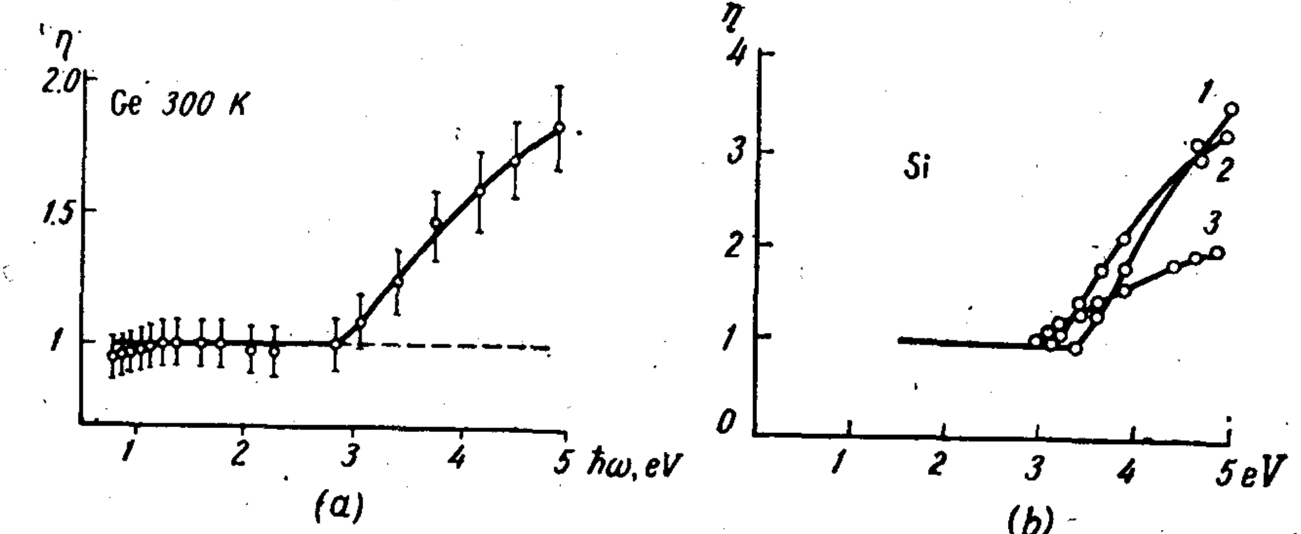


Fig. 132. Spectral dependence of quantum efficiency in germanium (a) and the effect of temperature on spectral quantum efficiency in silicon (b):
1—100 K; 2—300 K; 3—400 K

Considering the processes of the increase or the decrease in non-equilibrium carrier concentration from a steady-state value we demonstrated that their time dependence is exponential or tangensoidal (and hyperbolic).

However, in some cases the kinetics of the photoresistive effect takes a different course from those discussed above. The transient curves for the increase in excess carrier concentration are S-shaped (Fig. 133). The S-shape is the more pronounced (i.e., it takes longer for the steady-state concentration to be established) the smaller is the step of the light pulse. These facts may be explained with the aid of the *traps* concept.

Suppose the semiconductor contains traps whose concentration is M. They are able to trap particles of one type and retain them for a period of time θ. Consider two possible cases: $\theta \ll \tau_f$ and $\theta \gg \tau_f$. In the first case the particle may be repeatedly trapped (α-type trap), in the second—only once (β-type trap). The process of establishing a steady state may be described on the basis of the continuity equation which should be augmented by terms reflecting the particle exchange between the traps and the energy bands. Omitting the computations we shall confine ourselves to some qualitative considerations. If the traps are free, they will at first trap, for example, the electrons, thereby decreasing the carrier concentration growth rate and prolonging the transient state. In the same way the discharge of the traps will slow down the decrease process of carrier concentration from a steady-state level. The initial stage of the increase or decrease should be determined by the minimum characteristic time, i.e. by the trapping time θ in case of the α-type traps, and by the lifetime τ_f in case of the β-type traps. However, as the traps become filled their part in the transient process diminishes, and this explains the S-shape of the curves describing the time dependence of the carrier concentration increase to the steady state.

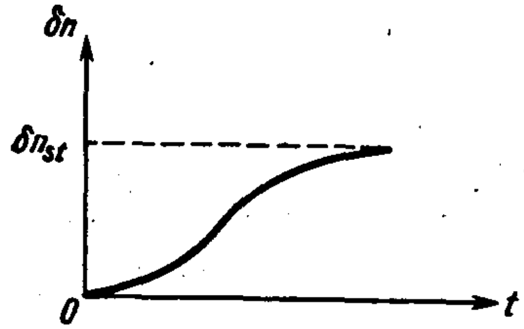


Fig. 133. S-shaped increase in the stationary concentration of non-equilibrium charge carriers

The existence of traps throws light, for example, on the part played by background illumination: the illumination of the material by long-wave light capable of photo-ionization of the traps enhances its photoconductivity.

The trap concept is at the present time very widely used in the theory of the photoresistive effect. Here is an example to illustrate the point. Suppose the semiconductor contains traps lying below the Fermi level. Suppose their trapping cross section for free holes is σ_p^*, and for free electrons, σ_n^*. In equilibrium the quasi-Fermi levels F_n^*, F_p^*, F_M^* coincide. In case of high donor concentrations the Fermi level lies near E_c, and the traps are filled with electrons. If the semiconductor is illuminated with light $\hbar\omega \geqslant \Delta E_0$, this will result in electron transitions from the valence band to the conduction band and in an equal increase in the electron and the hole concentrations. The deviation from thermodynamical equilibrium displaces the quasi-Fermi levels: F_n^* rises, and F_p^* drops. The displacement of F_M will be determined by the relation connecting σ_p^* and σ_n^*.

For $\sigma_p^* > \sigma_n^*$ the trapping probability of holes is greater than that of electrons, and F_M will drop until the equilibrium between the band and the discrete level is attained. The increase in electron concentration exceeds that of holes, and the photoconductivity is mainly of electron type (unipolar).

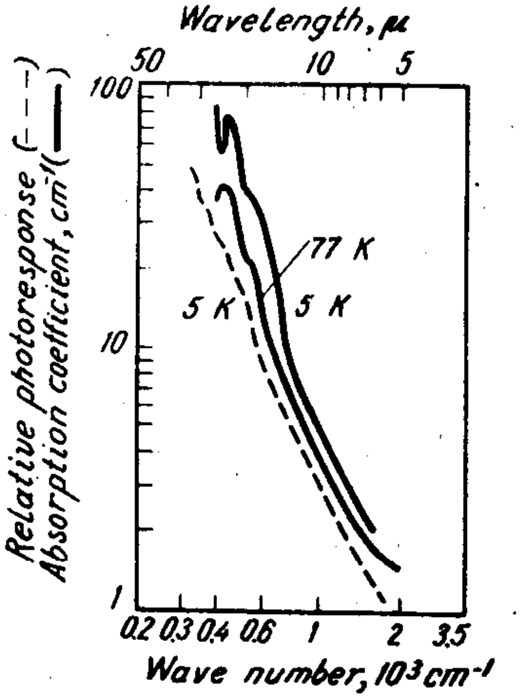


Fig. 134. Absorption coefficient and relative photoresponse for n-type antimony-doped germanium

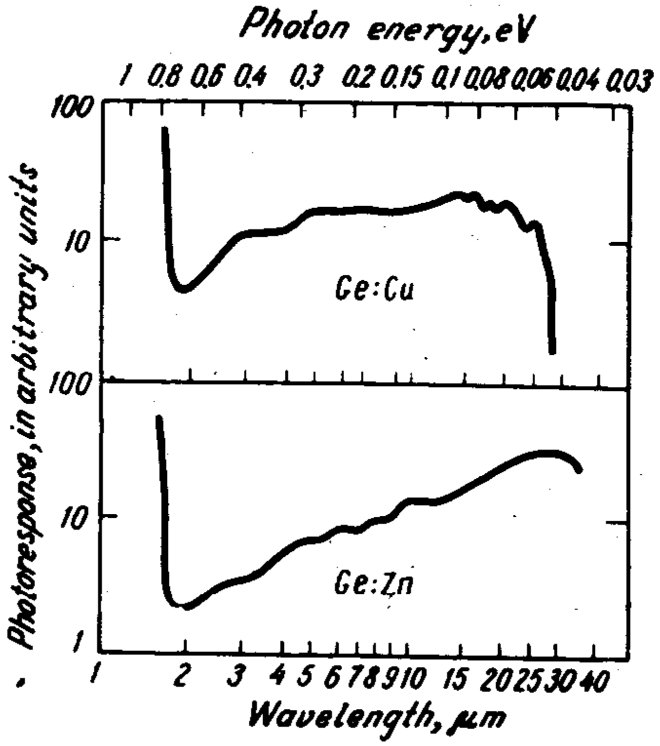


Fig. 135. Spectral dependence of photoresistive effect (in arbitrary units) in germanium doped with copper and zinc

For $\sigma_n^* > \sigma_p^*$ the Fermi level F_M rises, the traps are filled with electrons, and the hole concentration increase exceeds that of the electrons. Photoconductivity (possibly, full conductivity if $\sigma_{pc} > \sigma_d$) becomes hole-type. Such is the explanation of the unipolar photoresistive effect in conditions of intrinsic light absorption.

In the same way one may explain several other results in the kinetics of the photoresistive effect.

The spectral dependence of the photoconductivity is determined foremost by the spectral dependence of the carrier generation rate. The $\sigma_{pc}(\lambda)$ curves for the impurity internal photoeffect coincide quite well with the $\alpha(\lambda)$ absorption curves (Figs. 134 and 127). In some cases the $\sigma_{pc}(\lambda)$ spectrum consists of wider bands than the absorption bands (Fig. 135).

The spectral dependence of the photoresistive effect in the intrinsic absorption range differs from the intrinsic absorption spectrum. The position of the intrinsic photoresistive effect boundary coincides with that of the intrinsic absorption. However, as the photon energy increases, the spectral curve of the internal photoeffect decreases rapidly after passing through a maximum despite the fact that absorption in this region remains high (Fig. 136).

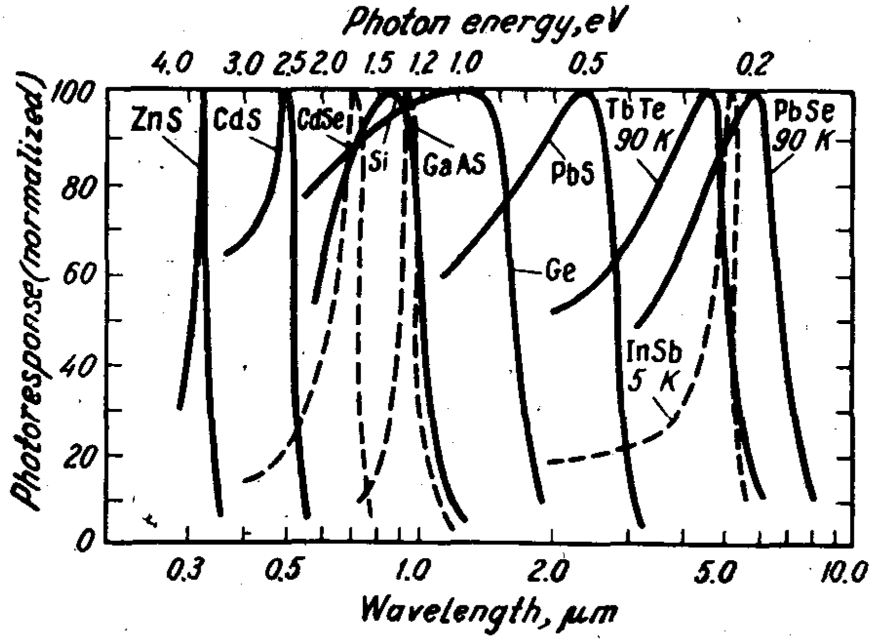


Fig. 136. Photoresistive effect spectra in the intrinsic absorption region

To understand the causes of this one should take account of the fact that owing to intense light absorption the carrier generation rate falls off drastically as light penetrates into the material:

$$G(x) = \eta \alpha q_0 e^{-\alpha x} = G(0)\, e^{-\alpha x}. \qquad (80.26)$$

For $x \gg \alpha^{-1} G(x) \cong 0$, and there is no increase in conductivity. This means that photoconductivity should depend on the sample geometry.

Characteristic of the material is the value of photosensitivity defined as the ratio of photoconductivity to light intensity:

$$S_{pc} = \frac{\sigma_{pc}}{J}. \qquad (80.27)$$

In the SI system S_{pc} is measured in ohm^{-1}m^{-1}/(W·m^{-2}) = = m/(ohm·W). The spectral characteristic of the photoresistive effect should be determined specifically as the dependence of photosensitivity on the frequency or on the wavelength of light. The photocurrent J_{pc} is sometimes termed photoresponse.

Summary of Sec. 80

1. The photoresistive, or the internal photoeffect, consists in the variation of the resistance (or the conductance) of a semiconductor sample caused by irradiation.

2. The variation of conductivity $\delta\sigma = \sigma_{pc}$ is termed photoconductivity. The ratio of photoconductivity to incident light intensity J is termed photosensitivity

$$S_{pc} = \frac{\sigma_{pc}}{J}. \tag{80.1s}$$

3. Photoconductivity is determined by the excess concentrations of minority and majority carriers:

$$\sigma_{pc} = e_n\mu_n\delta n + e_p\mu_p\delta p. \tag{80.2s}$$

In the steady state the excess carrier concentration is determined by the first characteristic relation

$$\delta n = G_n\tau_f^n = \eta_n\alpha q\tau_f^n, \tag{80.3s}$$

$$\delta p = G_p\tau_f^p = \eta_p\alpha q\tau_f^p, \tag{80.4s}$$

$$\sigma_{pc} = (e_n\mu_n\eta_n\tau_n + e_p\mu_p\eta_p\tau_p)\,\alpha q. \tag{80.5s}$$

4. The photocurrent is determined by the relation

$$I_{pc} = S\sigma_{pc}\,E = \left(\frac{\eta_n\tau_f^n}{t_n} + \frac{\eta_p\tau_f^p}{t_p}\right)e_pSl\alpha q, \tag{80.6s}$$

or

$$I_{pc} = e_pA'G, \tag{80.7s}$$

where $G = Sl\alpha q$ is the total number of photons absorbed in the volume Sl of the semiconductor sample, $A' = \frac{\eta_n\tau_f^n}{t_n} + \frac{\eta_p\tau_f^p}{t_p}$ is the "amplification factor", and t_n and t_p is the electron and hole transit time determined by the condition $\mu_n Et_n = \mu_p Et_p = l$.

The relation (80.7s) bears the name of the second characteristic relation of the photoresistive effect.

5. The dependence of photoconductivity on light frequency is determined by the absorption spectrum $\alpha(\omega)$ or indirectly by the carrier lifetime.

6. To explain the kinetics of the photoresistive effect use is made of the concept of traps of varying parameters.

7. The photoresistive effect is utilised in photoelectric devices such as photoresistors capable of converting radiation signals into electric signals.

81. DEMBER EFFECT. PHOTOVOLTAIC EFFECT

Consider a homogeneous semiconductor illuminated with light. As light penetrates into the semiconductor its intensity abates in compliance with the Bouguer-Lambert law

$$J(x) = J(0)\,e^{-\alpha x}, \quad q(x) = q(0)\,e^{-\alpha x}. \tag{81.1}$$

Absorption of light results in charge carrier generation which declines exponentially with the distance inside the semiconductor:

$$G(x) = G(0)\,e^{-\alpha x}; \quad G(0) = \eta\alpha q(0). \tag{81.2}$$

Non-uniform generation causes the particles to diffuse into the semiconductor. But since the diffusion coefficients of electrons and holes are different, the mobile charge carriers are separated. The electrons whose mobility is as a rule greater penetrate deeper into the semiconductor than the holes. The illuminated surface acquires a positive charge, and the dark surface—a negative charge so that an electric field directed along the light beam is established.

The appearance of an electric field inside a homogeneous semiconductor illuminated with light is termed Dember, or *crystal photoeffect*. Find the field intensity of the Dember effect using the kinetic equation which in the form we have been using is valid for steady-state conditions. To suit our end the equation should have its equilibrium kinetic coefficients and other parameters replaced by their non-equilibrium values.

Assume the semiconductor to be homogeneous, in isothermal conditions, with $\mathbf{B} = 0$. Write the equation for current density

$$\mathbf{j} = \sum_\alpha e_\alpha K_{11(\alpha)}\mathbf{E} - \sum_\alpha e_\alpha K_{11(\alpha)} \nabla F_\alpha. \tag{81.3}$$

The Dember effect is observed in a semiconductor in the absence of a current ($\mathbf{j} = 0$). Therefore

$$\mathbf{E} = \frac{\sum\limits_\alpha e_\alpha K_{11(\alpha)} \nabla F_\alpha}{\sum\limits_\alpha e_\alpha^2 K_{11(\alpha)}}. \tag{81.4}$$

Suppose the semiconductor contains electrons ($\alpha = 1$) and holes ($\alpha = 2$). Then

$$e_1^2 K_{11(1)} = e_n \mu_n n = \sigma_n; \quad e_2^2 K_{11(2)} = e_p \mu_p p = \sigma_p. \tag{81.5}$$

Besides, according to (66.5)

$$\nabla F_\alpha = kT\,\frac{\nabla n_\alpha}{n_\alpha}, \tag{81.6}$$

therefore

$$\mathbf{E} = \frac{\sum\limits_\alpha e_\alpha \frac{n_\alpha\langle\tau_\alpha\rangle}{m_\alpha^*}\frac{\nabla n_\alpha}{n_\alpha} kT}{\sigma_n + \sigma_p} = \frac{\sum\limits_\alpha \mu_\alpha \nabla n_\alpha kT}{\sigma_n + \sigma_p} = \frac{\sum\limits_\alpha e_\alpha D_\alpha \nabla n_\alpha}{\sigma} = \frac{-\sum\limits_\alpha \mathbf{j}_{D_\alpha}}{\sum\limits_\alpha \sigma_\alpha} = -\frac{\mathbf{j}_{D_n} + \mathbf{j}_{D_p}}{\sigma_n + \sigma_p}. \tag{81.7}$$

The equation (81.7) shows that non-uniform illumination causes diffusion fluxes of electrons and holes to flow in one direction — into the semiconductor, the directions of the diffusion currents $\mathbf{j}_{Dn}$ and $\mathbf{j}_{Dp}$ being opposite. The resulting separation of charges establishes an electric field **E** creating a drift current which compensates the diffusion current. The electric field **E** is determined by the difference between the diffusion fluxes of the electrons and holes.

Setting the potential of the dark side to be zero ($\varphi(\infty)=0$) find the potential of the illuminated side $\varphi(0)$:

$$-\int_0^\infty (\mathbf{E}\,d\mathbf{l}) = \varphi(\infty) - \varphi(0) = -\varphi(0), \tag{81.8}$$

$$\varphi(0) = \int_0^\infty \frac{e_n D_n \dfrac{dn}{dx} + e_p D_p \dfrac{dp}{dx}}{\sigma}\,dx. \tag{81.9}$$

To calculate this integral $\sigma(x)$ should be known, since owing to diffusion into the volume of non-equilibrium charge carriers the volume conductivity varies with the depth x. Consider a feeble generation for which $\sigma \cong \sigma_0$. In this case the calculation of the integral presents no difficulties:

$$\varphi(0) = \frac{e_n D_n}{\sigma_0}\int_0^\infty dn + \frac{e_p D_p}{\sigma_0}\int_0^\infty dp = \frac{-e_n D_n\,\delta n(0) - e_p D_p\,\delta p(0)}{\sigma}. \tag{81.10}$$

If intrinsic absorption is responsible for generation, $\delta n(0) = \delta p(0) = \alpha q \eta$, and

$$\varphi(0) = \mathscr{E}^D = \frac{e_p\,\delta p(0)}{\sigma}(D_n - D_p) = -\frac{kT\,\delta p(0)}{\sigma}(\mu_n + \mu_p) =$$
$$= \frac{kT}{\sigma}(|\mu_n| - \mu_p)\,\alpha q \eta. \tag{81.11}$$

The e.m.f. of the Dember effect $\mathscr{E}^D$ is determined by the difference between the mobility moduli, i.e. $\mathscr{E}^D = 0$ when $|\mu_n| = \mu_p$; it is proportional to the intensity of light J, the proportionality being retained as long as the assumption $\sigma \cong \sigma_0$, remains valid. For high intensities there will be a rise in σ and a slow down in the increase in $\mathscr{E}^D$.

In connection with the Dember effect one peculiar feature of non-equilibrium carrier diffusion should be noted. The direction of the Dember field is such that it hinders the diffusion of the more mobile carriers and enhances the diffusion of the less mobile.

The combined diffusion of electrons and holes is termed bipolar diffusion. It may be described with the aid of the continuity equation in the same way as it was done in the case of unipolar diffusion in Sec. 66. To this end substitute the expressions for the currents $\mathbf{j}_n$ and $\mathbf{j}_p$ into the equations (80.3) and (80.4). We obtain for a steady state

$$D_n \nabla^2 n + \mu_n n \operatorname{div} \mathbf{E} + \mu_n (\nabla n \mathbf{E}) + G_n - \frac{n-n_0}{\tau_f^n} = 0, \qquad (81.12)$$

and

$$D_p \nabla^2 p + \mu_p p \operatorname{div} \mathbf{E} + \mu_p (\nabla p \mathbf{E}) + G_p - \frac{p-p_0}{\tau_f^p} = 0. \qquad (81.13)$$

The equations (81.12) and (81.13) are connected by means of the field $\mathbf{E}$:

$$\operatorname{div} \mathbf{E} = -\frac{4\pi}{\varepsilon} \rho = -\frac{4\pi}{\varepsilon} [e_n \,\delta n + e_p \,\delta p]. \qquad (81.14)$$

Consider the solution of the equations (81.12) and (81.13) for one specific but nevertheless important case—electric neutrality is maintained in the volume: $\rho = 0$, $\delta n = \delta p$, and therefore $\operatorname{div} \mathbf{E} = 0$. Since $\delta n = n - n_0 = \delta p = p - p_0$, $\tau_f^n = \tau_f^p = \tau_f$ i.e. the lifetimes of electrons, holes, and carrier pairs are equal.

Multiplying (81.12) by σ_{p0} and (81.13) by σ_{n0}, and adding them we obtain

$$(\sigma_{p0} D_n + \sigma_{n0} D_p) \nabla^2 n + (\mu_n \sigma_{p0} + \mu_p \sigma_{n0}) (\mathbf{E} \nabla n) + G_n \sigma_{p0} + G_p \sigma_{n0} - \\ - \frac{n-n_0}{\tau_f} (\sigma_{p0} + \sigma_{n0}) = 0, \qquad (81.15)$$

or in the unidimensional case

$$\frac{D_n \sigma_{p0} + D_p \sigma_{n0}}{\sigma_{p0} + \sigma_{n0}} \frac{d^2 (n-n_0)}{dx^2} + \frac{(\mu_n \sigma_{p0} + \mu_p \sigma_{n0})}{\sigma_{p0} + \sigma_{n0}} \mathbf{E} \frac{d (n-n_0)}{dx} + \\ + \frac{G_n \sigma_{p0} + G_p \sigma_{n0}}{\sigma_{p0} + \sigma_{n0}} - \frac{n-n_0}{\tau_f} = 0. \qquad (81.16)$$

Introduce the notations:
bipolar diffusion coefficient

$$D = \frac{D_n \sigma_{p0} + D_p \sigma_{n0}}{\sigma_{p0} + \sigma_{n0}}; \qquad (81.17)$$

bipolar drift mobility

$$\mu_E = \frac{\mu_n \sigma_{p0} + \mu_p \sigma_{n0}}{\sigma_{p0} + \sigma_{n0}}. \qquad (81.18)$$

μ_E and D, as may be seen from (81.17) and (81.18), do not satisfy the Einstein relation. A new parameter may be introduced connected with D by the Einstein relation—the bipolar diffusion mobility

$$\mu_D = \frac{e_p}{kT} D = \frac{\mu_p \sigma_{n0} - \mu_n \sigma_{p0}}{\sigma_{p0} + \sigma_{n0}}. \qquad (81.19)$$

An important property of the parameters D, μ_E and μ_D is their dependence on the relation between the concentrations n_0 and p_0. In an intrinsic semiconductor ($n_0 = p_0$) the parameters are

$$D = 2\frac{D_p D_n}{D_p + D_n}; \quad \frac{1}{D} = \frac{1}{2}\left(\frac{1}{D_p} + \frac{1}{D_n}\right); \qquad (81.20)$$

$$\mu_E = 0; \qquad (81.21)$$

$$\mu_D = 2\frac{\mu_p |\mu_n|}{\mu_p + |\mu_n|}; \quad \frac{1}{\mu_D} = \frac{1}{2}\left(\frac{1}{\mu_p} + \frac{1}{|\mu_n|}\right). \qquad (81.22)$$

In the extrinsic conductivity range, for instance when $\sigma_{p0} \ll \sigma_{n0}$,

$$D = D_p, \quad \mu_E = \mu_p, \quad \mu_D = \mu_p; \qquad (81.23)$$

when $\sigma_{n0} \ll \sigma_{p0}$

$$D = D_n, \quad \mu_E = \mu_n, \quad \mu_D = -\mu_n. \qquad (81.24)$$

Thus, *in the extrinsic conductivity range the parameters D and μ_E coincide with the corresponding parameters of minority carriers, μ_D being equal to the modulus of the minority carrier mobility.*

Making use of the notations (81.17) and (81.18) re-write (81.16):

$$D\frac{d^2(n-n_0)}{dx^2} + \mu_E E\frac{d(n-n_0)}{dx} - \frac{n-n_0}{\tau_f} = \text{G}, \qquad (81.25)$$

where

$$\text{G} = \frac{\text{G}_n \sigma_{p0} + \text{G}_p \sigma_{n0}}{\sigma_{p0} + \sigma_{n0}} \qquad (81.26)$$

is the "bipolar" generation rate.

The equation (81.25) for $\text{G} = 0$ coincides with the equation (66.15). Denoting

$$L^2 = D\tau_f, \qquad (81.27)$$

$$\frac{\mu_E \text{E}}{D} = \frac{\mu_E \text{E} \tau_f}{D\tau_f} = \frac{l_E}{L^2}, \qquad (81.28)$$

we obtain for (81.25)

$$\frac{d^2(n-n_0)}{dx^2} + \frac{l}{L^2}\frac{d(n-n_0)}{dx} - \frac{n-n_0}{L^2} = -\frac{\text{G}}{D}. \qquad (81.29)$$

The general solution of the homogeneous equation may be presented in the form

$$(n-n_0) = Ae^{-\frac{x}{l}} + Be^{\frac{x}{l}}. \qquad (81.30)$$

The partial solution may be found if $G(x)$ is known. If the charge carriers are generated by light,

$$G(x) = G(0)\, e^{-\alpha x}. \tag{81.31}$$

The partial solution for this case is

$$(n - n_0) = -\frac{G(0)\, e^{-\alpha x}}{D\left[\alpha^2 - \frac{\alpha l_E}{L^2} - \frac{1}{L^2}\right]}, \tag{81.32}$$

which may be checked by directly substituting (81.31) into (81.29).

For most cases of intrinsic absorption it may be assumed that $G(x) \cong 0\ (x \neq 0)$. The coefficients A and B of the general solution may be found from the boundary conditions for concentrations and currents provided the surface recombination currents are taken into account.

We shall omit these unsophisticated but tedious computations. Note one peculiarity of the $[n - n_0](x)$ distribution. *For weak fields* E *the non-equilibrium carrier concentration distribution is determined by the diffusion length L. For strong fields in extrinsic semiconductors the drag distance is equal to the drift distance.* But as the conductivity of the semiconductor approaches the intrinsic value, the drag distance tends toward the diffusion length. In an intrinsic semiconductor the electric field no matter how strong (provided it does not violate the electric neutrality) does not affect the distribution of excess carrier concentration.

Consider now the qualitative aspects of the photogalvanic effect.

It was demonstrated in Sec. 72 that an inhomogeneous semiconductor contains built-in fields $\mathbf{E}^i$ and space charges the values of which may be found from the relation (72.4) which we shall re-write as follows:

$$\mathbf{E}^i = -\frac{\mathbf{j}_{Dn0} + \mathbf{j}_{Dp0}}{\sigma_{p0} + \sigma_{n0}} = \frac{D_p}{\mu_p}\,\frac{\nabla p_0 - b\nabla n_0}{p_0 + bn_0}. \tag{81.33}$$

The expression (81.33) exactly coincides with (81.7), but there is a difference in principle. *The field* $\mathbf{E}^i$ *is a thermodynamically equilibrium field that compensates for the diffusion current and provides for the total current in the semiconductor volume to be zero. For this reason the field* $\mathbf{E}^i$ *cannot produce a current in an external circuit while the Dember field, being a non-equilibrium field, produces such a current.*

Suppose light is falling on an inhomogeneous semiconductor. The electron and the hole generated by light are acted upon by the built-in field $\mathbf{E}^i$ which makes them drift. But while diffusion makes them move in the same direction, the field $\mathbf{E}^i$ makes them move *in opposite directions*, the electrons moving against the field, and the holes—along it.

The separation of non-equilibrium carriers causes *a non-equilibrium electric field* **E*** *directed against the field* **E**^{i} to be established. Hence, the non-equilibrium built-in field **E*** prevents the non-equilibrium carriers from being separated. Obviously, in the limiting case $\delta n = \delta p \rightarrow \infty$ the volume fields **E*** and **E**^{i} will completely compensate each other, the field and the space charge will turn zero, the potential barriers which existed in the volume of the semiconductor will be removed, and the semiconductor will become uniform and intrinsic $\delta n = n = \delta p = p$ (but it will remain a non-equilibrium one). A potential difference with the modulus equal to the ratio of the barrier height to the electron charge will be established between the points of the semiconductor which were formerly separated by potential barriers. Hence, *when non-equilibrium charge carriers are generated in the area containing built-in fields these fields result in the establishment of a potential difference between any two points located in the area.* The non-equilibrium carriers may also be in evidence if they are generated at distances from the built-in fields **E**^{i} less than the diffusion length. In this case they will diffuse to the area of the built-in field which will separate them.

Consider one important case of carrier generation in the *p-n* junction region containing a space charge and a built-in field directed from the *n*- to the *p*-region. The holes and electrons generated in the *n*- and *p*-regions within the distances from the space charge region of the order of L_p and L_n, respectively, will be separated by the field **E**$_k$ so that non-equilibrium electrons will be transported to the *n*-region, and the holes, to the *p*-region, i.e. *the contact field moves the minority carriers across the space-charge region.* The electron-type region acquires a negative charge, and the hole-type region—a positive. Should such a semiconductor be connected to a closed circuit a current would be initiated in it directed from the *p*-region to the *n*-region and causing the disappearance of the excess carrier concentration. The mechanism by which the fields **E*** are established is such that the maximum ph.e.m.f. across the *p*-n junction cannot exceed the value of φ^k which determines the height of the potential barrier. This phenomenon provides the basis for the operation of *p-n* junction photodiodes and convertors of luminous energy into electric energy ("solar batteries"). The photovoltaic effect is observed in all illuminated rectifying contacts and may be used to study the properties of such contacts. Another application is *the detection of inhomogeneities of a semiconductor by the light-probe method.*

Figure 137 shows the spectrum S_λ of the photovoltaic effect in cadmium telluride. The long-wave photosensitivity bands are outside the intrinsic absorption region. This means that the photovoltaic effect may also take place in the case of unipolar minority carrier generation caused by optical or thermo-optical transitions. A peak corresponding to exciton absorption may be observed on the curve S_λ. The only possible cause of its appearance is the dissociation of the exciton.

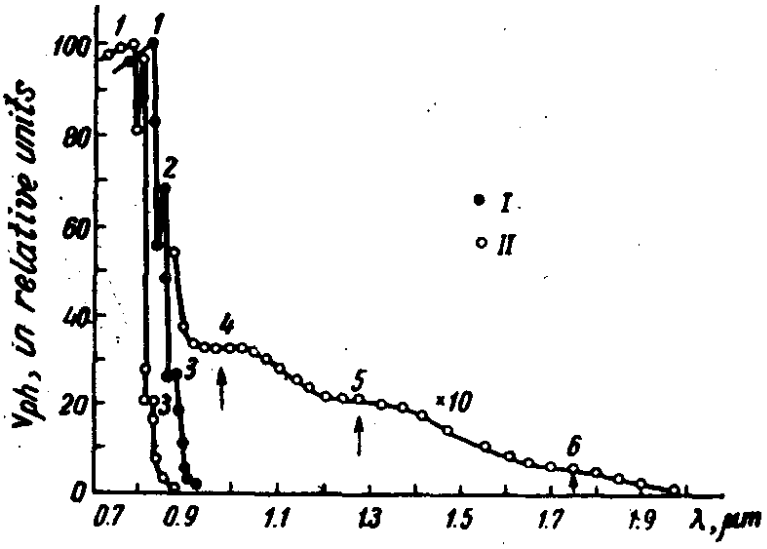


Fig. 137. Photovoltaic effect spectrum in cadmium telluride at room (I) and at liquid nitrogen temperature (II)

The photo-electromotive force in solid solutions of variable composition and, therefore, of a variable forbidden band width (variable-band crystals) may also be classified as a photovoltaic effect.

In the solid solution of $Cd_xHg_{1-x}Te$ the forbidden band width can vary at room temperature from 1.5 eV for cadmium telluride ($x=1$) to 0 in mercury telluride ($x=0$). Since the energy extrema in this case are a function of the co-ordinate, an electric field is established which acts on conduction electrons and holes:

$$E_n = -\frac{1}{e_n}\frac{dE_c(x)}{dx}, \qquad (81.34)$$

$$E_p = -\frac{1}{e_p}\frac{dE_v(x)}{dx}. \qquad (81.35)$$

These fields differ in magnitude and in sign. The forces acting upon electrons and holes act in the same direction: from the region of smaller forbidden band width to the region of greater forbidden band width. In conditions of thermodynamical equilibrium these fields do not cause an electromotive force to be established in an external circuit. But should the equilibrium be violated, for instance by the absorption of light, the charge carriers would begin to flow in the direction of the narrow-band region. Since normally the mobility of electrons is greater than that of holes, their velocity exceeds the velocity of holes with the result that the charges are separated in space. This causes a non-equilibrium electric field to be established which, in turn, produces an electric current in an external circuit. The photosensitivity spectrum of the variable-band crystals is quite wide: from the visible region to 10-20 μm.

Summary of Sec. 81

1. Simultaneous diffusion of non-equilibrium charge carriers of opposite signs is termed bipolar. It is characterized by the bipolar diffusion coefficient

$$D=\frac{D_n\sigma_{p0}+D_p\sigma_{n0}}{\sigma_{p0}+\sigma_{n0}}, \qquad (81.1s)$$

where $\sigma_{p0}+\sigma_{n0}$ is the equilibrium conductivity. The distribution of the charge carriers is determined by the drag distance which is practically equal to the bipolar diffusion length defined by the equation

$$L^2=D\tau_f=\frac{L_n^2\sigma_{p0}+L_p^2\sigma_{n0}}{\sigma_{p0}+\sigma_{n0}}, \qquad (81.2s)$$

which is valid for weak fields E, or for strong fields in a nearly intrinsic semiconductor.

2. The distribution of non-equilibrium concentration is determined by the equation

$$n(x)-n_0=p-p_0=Ae^{-\frac{x}{l}}+Be^{\frac{x}{l}}-\frac{G(0)e^{-\alpha x}}{D\left[\alpha^2-\frac{\alpha l_E}{L^2}-\frac{1}{L^2}\right]}. \qquad (81.3s)$$

The coefficients *A* and *B* may be determined from the boundary conditions.

3. The Dember, or the crystal, photoeffect consists in the appearance of an electric field directed along the beam of intensely absorbed light. The field intensity or the e.m.f. $\mathcal{E}^D$ is proportional to the difference between the electron and hole diffusion coefficients because the Dember field compensates the difference between electron and hole diffusion currents. If the diffusion coefficients are equal, the currents will be equal in magnitude and opposite in direction and there will be no Dember field.

4. The separation of non-equilibrium charge carriers in built-in fields establishes non-equilibrium fields in the volume of the semiconductor capable of producing a current in an external circuit. The photovoltaic effect consists in the appearance of an e.m.f. across a *p-n* junction or a rectifying metal-semiconductor contact when non-equilibrium carriers are generated by light in their region.

82. PHOTOMAGNETOELECTRIC EFFECT

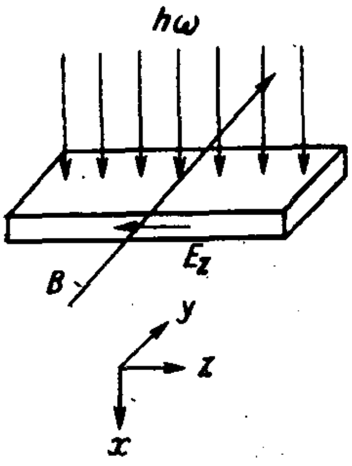


Fig. 138. A set-up for the observation of the Kikoin-Noskov effect

An electric field is established in a semiconductor placed in a magnetic field and illuminated with light which it intensely absorbs. This phenomenon is termed Kikoin-Noskov, or photomagnetoelectric, effect (PhMEE). The cause of the electric field is the deviation of the drift and diffusion carrier flows from their original direction by the magnetic field. In experimental studies of the photomagnetoelectric effect one of the faces of a semiconductor bar is illuminated with light, the magnetic field being directed so as to be perpendicular both to the direction of the light flux and to the other pair of faces. The electric field is then observed in the direction perpendicular to the magnetic field and to the light flux (Fig. 138). The Kikoin-Noskov effect is similar to two phenomena: to the Hall effect, since there is a drift current in the semiconductor, and to the Nernst-Ettingshausen effect, since a diffusion current is initiated in a non-uniformly illuminated semiconductor.

The photomagnetoelectric effect in steady-state conditions may be described on the basis of the kinetic equation. Accounting for the action of the electric and magnetic fields and of the chemical potential gradient write

$$\mathbf{j}=\sum_{\alpha} e_{\alpha}^{2}K'_{11\,(\alpha)}\mathbf{E}-\sum_{\alpha} e_{\alpha}K'_{11\,(\alpha)}\,\nabla F_{\alpha}+$$

$$+\left[\left(\sum_{\alpha}\frac{e_{\alpha}^{3}}{\overset{*}{m}_{\alpha}}K'_{12\,(\alpha)}\mathbf{E}-\sum_{\alpha}\frac{e_{\alpha}^{2}}{\overset{*}{m}_{\alpha}}K'_{12\,(\alpha)}\nabla F_{\alpha}\right),\ \mathbf{B}\right]. \qquad (82.1)$$

The first term describes the drift current, the second—the diffusion current, the third—the Hall galvanomagnetic current, and the fourth—the magnetodiffusion current. The expression (82.) takes into account currents produced by charge carriers of different types. Transform the terms in (82.1) as follows:

$$e_{\alpha}^{2}K'_{11\,(\alpha)}=e_{\alpha}\mu'_{\alpha}n_{\alpha}=\sigma'_{\alpha}; \qquad (82.2)$$

$$e_{\alpha}K'_{11\,(\alpha)}\,\nabla F_{\alpha}=\mu'_{\alpha}n_{\alpha}kT\,\frac{\nabla n_{\alpha}}{n_{\alpha}}=e_{\alpha}D'_{\alpha}\nabla n_{\alpha}; \qquad (82.3)$$

$$\frac{e_{\alpha}}{\overset{*}{m}_{\alpha}}K'_{12\,(\alpha)}=\frac{e_{\alpha}^{3}n_{\alpha}\langle\tau_{\alpha}'^{2}\rangle}{\overset{*}{m}{}_{\alpha}^{2}}=\frac{e_{\alpha}\langle\tau_{\alpha}'^{2}\rangle}{\overset{*}{m}_{\alpha}\langle\tau'_{\alpha}\rangle}\,\frac{e_{\alpha}\langle\tau'_{\alpha}\rangle}{\overset{*}{m}_{\alpha}}\,e_{\alpha}n_{\alpha}=\mu_{\alpha}^{\mathrm{H}}\sigma'_{\alpha}; \qquad (82.4)$$

$$\frac{e_{\alpha}}{\overset{*}{m}_{\alpha}}K'_{12\,(\alpha)}\nabla F_{\alpha}=\frac{e_{\alpha}^{2}}{\overset{*}{m}_{\alpha}}\,\frac{n_{\alpha}\langle\tau_{\alpha}'^{2}\rangle}{\overset{*}{m}_{\alpha}}\,kT\,\frac{\nabla n_{\alpha}}{n_{\alpha}}=$$

$$=\mu_{\alpha}^{\mathrm{H}}\mu'_{\alpha}kT\,\nabla n_{\alpha}=e_{\alpha}D'_{\alpha}\nabla n_{\alpha}\mu_{\alpha}^{\mathrm{H}}. \qquad (82.5)$$

All the quantities in (82.2-5) relate to non-equilibrium concentrations, and the stroke accounts for $(1+\mu_\alpha^2 B^2)$ in the denominator of the averaged relaxation times. In weak magnetic fields the term $\mu_\alpha^2 B^2$ may be neglected which will enable us to neglect the magnetic field dependence of the respective quantities.

Re-write (82.1) making use of (82.2-5):

$$\mathbf{j}=\Big(\sum_\alpha \sigma_\alpha\Big)\mathbf{E}-\Big(\sum_\alpha e_\alpha D_\alpha \nabla n_\alpha\Big)+\Big[\Big(\sum_\alpha \mu_\alpha^{\mathrm{H}}\sigma_\alpha \mathbf{E}-\sum_\alpha e_\alpha D_\alpha \nabla n_\alpha \mu_\alpha^{\mathrm{H}}\Big),\mathbf{B}\Big]=$$
$$=\sum_\alpha \mathbf{j}_{E\alpha}+\sum_\alpha \mathbf{j}_{D\alpha}+\Big[\sum_\alpha \mu_\alpha^{\mathrm{H}}\mathbf{j}_{E\alpha},\ \mathbf{B}\Big]+\Big[\sum_\alpha \mu_\alpha^{\mathrm{H}}\mathbf{j}_{D\alpha},\ \mathbf{B}\Big]. \qquad (82.6)$$

Expression (82.6) shows quite clearly the composition of the current which includes the Hall current, the diffusion current, and the magnetodiffusion current. Re-write (82.6) in co-ordinate projections making use of the fact that $\mathbf{B}=(0,\ B,\ 0)$:

$$j_x=\sigma E_x-\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha-\Big(\sum_\alpha \mu_\alpha^{\mathrm{H}}\sigma_\alpha\Big)E_z B+\Big(\sum_\alpha e_\alpha D_\alpha \nabla_z n_\alpha \mu_\alpha^{\mathrm{H}}\Big)B; \qquad (82.7)$$

$$j_y=\sigma E_y-\sum_\alpha e_\alpha D_\alpha \nabla_y n_\alpha; \qquad (82.8)$$

$$j_z=\sigma E_z-\sum_\alpha e_\alpha D_\alpha \nabla_z n_\alpha+\Big(\sum_\alpha \mu_\alpha^{\mathrm{H}}\sigma_\alpha\Big)E_x B-\Big(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_\alpha^{\mathrm{H}}\Big)B. \qquad (82.9)$$

Express E_x and E_z by solving the equation system (82.7) and (82.9). The system determinant Δ is

$$\Delta=\begin{vmatrix} \sigma & \Big(-\sum_\alpha \mu_\alpha^{\mathrm{H}}\sigma_\alpha\Big)B \\ \Big(\sum_\alpha \mu_\alpha^{\mathrm{H}}\sigma_\alpha\Big)B & \sigma \end{vmatrix}=\sigma^2+\Big(\sum_\alpha \sigma_\alpha \mu_\alpha^{\mathrm{H}}\Big)^2 B^2. \qquad (82.10)$$

For weak magnetic fields $(\mu_\alpha^2 B^2 \ll 1)$

$$\Delta=\sigma^2. \qquad (82.11)$$

For E_x and E_z we obtain the equations

$$E_x=\frac{1}{\Delta}\Big\{\sigma\Big[j_x+\Big(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha\Big)-\Big(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_\alpha^{\mathrm{H}}\Big)B\Big]+$$
$$+\Big(\sum_\alpha \mu_\alpha^{\mathrm{H}}\sigma_\alpha\Big)B\Big[j_z+\Big(\sum_\alpha e_\alpha D_\alpha \nabla_z n_\alpha\Big)+\Big(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_\alpha^{\mathrm{H}}\Big)B\Big]\Big\}; \qquad (82.12)$$

$$E_z=\frac{1}{\Delta}\Big\{\sigma\Big[j_z+\Big(\sum_\alpha e_\alpha D_\alpha \nabla_z n_\alpha\Big)+\Big(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_\alpha^{\mathrm{H}}\Big)B\Big]-$$
$$-\Big(\sum_\alpha \mu_\alpha^{\mathrm{H}}\sigma_\alpha\Big)B\Big[j_x+\Big(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha\Big)-\Big(\sum_\alpha e_\alpha D_\alpha \nabla_z n_\alpha \mu_\alpha^{\mathrm{H}}\Big)B\Big]\Big\}. \qquad (82.13)$$

For $B=0$ the expressions (82.12) and (82.13) take the form

$$E_x = \frac{j_x + \sum\limits_\alpha e_\alpha D_\alpha \nabla_x a_\alpha}{\sigma}, \tag{82.14}$$

$$E_z = \frac{j_z + \sum\limits_\alpha e_\alpha D_\alpha \nabla_z n_\alpha}{\sigma}, \tag{82.15}$$

i.e. E_x is the field of the ohmic voltage drop and E_z is the field that compensates for the diffusion fluxes. For $j_x = j_z = 0$ E_x turns into the Dember field, and E_z turns zero since $\nabla_z n_\alpha = 0$.

Consider the equations (82.12) and (82.13) in the assumption that the magnetic field does not affect the carrier distribution along the z-axis i.e. that $\nabla_z n_\alpha = 0$. Neglecting the terms containing B^2 we write

$$E_x = \frac{1}{\sigma^2}\left\{\sigma\left[j_x + \left(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha\right)\right] + \left(\sum_\alpha \mu_\alpha^{\mathrm{H}} \sigma_\alpha\right) j_z B\right\}, \tag{82.16}$$

$$\begin{aligned} E_z = \frac{1}{\sigma^2}\Bigg\{\sigma\left[j_x + \left(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_\alpha^H\right) B\right] - \\ -\left(\sum_\alpha \mu_\alpha^{\mathrm{H}} \sigma_\alpha\right) B \left[j_x + \sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha\right]\Bigg\}. \end{aligned} \tag{82.17}$$

The expression (82.17) describes the Kikoin-Noskov effect for different conditions of observation. In conditions of experiment $j_x = 0$. Setting $j_z = 0$ we arrive at the equation for the fields E_z, or at the corresponding potential difference:

$$E_z = B\,\frac{\left(\sum\limits_\alpha \sigma_\alpha\right)\left(\sum\limits_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_\alpha^{\mathrm{H}}\right) - \left(\sum\limits_\alpha \mu_\alpha^H \sigma_\alpha\right)\left(\sum\limits_\alpha e_\alpha D_\alpha \nabla_x n_\alpha\right)}{\left(\sum\limits_\alpha \sigma_\alpha\right)^2}. \tag{82.18}$$

In short-circuit conditions $E_z = 0$, and from (82.17) we obtain the expression for the current density j_z:

$$-j_z = B\,\frac{\left(\sum\limits_\alpha \sigma_\alpha\right)\left(\sum\limits_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_\alpha^{\mathrm{H}}\right) - \left(\sum\limits_\alpha \mu_\alpha^{\mathrm{H}} \sigma_\alpha\right)\left(\sum\limits_\alpha e_\alpha D_\alpha \nabla_x n_\alpha\right)}{\sigma}. \tag{82.19}$$

Comparing (82.18) and (82.19) we note that E_z measured in open-circuit conditions and j_z measured in short-circuit conditions are related by means of an obvious expression

$$-j_z^{sc} = \sigma E_z^{oc}. \tag{82.20}$$

For the sake of generality consider the condition of "double short-circuit" in which $E_z = E_x = 0$. Express j_x from (82.16)

$$-j_x = \left(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha\right) + \frac{\left(\sum_\alpha \mu_\alpha^{\mathrm{H}} \sigma_\alpha\right) B j_z}{\sigma} \qquad (82.21)$$

and substitute it into (82.17):

$$0 = \sigma j_z + \sigma\left(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_\alpha^{\mathrm{H}}\right) B - \left(\sum_\alpha \mu_\alpha^{\mathrm{H}} \sigma_\alpha\right)\left(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha\right) B +$$
$$+ \left(\sum_\alpha \mu_\alpha^{\mathrm{H}} \sigma_\alpha\right) B \left(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha\right) = \sigma j_x + \sigma\left(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_\alpha^{\mathrm{H}}\right) B \quad (82.22)$$

or

$$j_x = -\left(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_\alpha^{\mathrm{H}}\right) B. \qquad (82.23)$$

E_z may be found in conditions of the "second short-circuit" ($E_x = 0$) when $j_z = 0$.

From (82.16) we obtain

$$j_x = -\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha, \qquad (82.24)$$

and from (82.17) making use of (82.24)

$$E_z = \frac{\left(\sum_\alpha e_\alpha D_\alpha \nabla_x n_\alpha \mu_p^{\mathrm{H}}\right) B}{\sigma}, \qquad (82.25)$$

in full agreement with (82.23).

The equations (82.16) and (82.17) describe a more general case than those considered above since they allow for a current j_z to be passed through the sample and for E_z to be measured simultaneously.

Write the expression for E_z in conditions of $j_x = j_z = 0$ taking account of electron and hole generation:

$$E_z = \frac{e_n D_n \nabla_x n \mu_n^{\mathrm{H}} + e_p D_p \nabla_x p \mu_p^{\mathrm{H}}}{\sigma} B - \frac{\left(\sum_\alpha \mu_\alpha^{\mathrm{H}} \sigma_\alpha\right)(e_n D_n \nabla_x n + e_p D_p \nabla_x p)}{\sigma^2} B =$$
$$= \frac{e_p \nabla_x p\,(D_p \mu_p^{\mathrm{H}} - D_n \mu_n^{\mathrm{H}})}{\sigma} B - \frac{(\sigma_p \mu_p^{\mathrm{H}} + \sigma_n \mu_n^{\mathrm{H}})\, e_p \nabla_x p\,(D_p - D_n)}{\sigma^2} B. \quad (82.26)$$

In the impurity absorption range unipolar diffusion takes place, and for $\nabla_x n = 0$ we obtain

$$E_z = \frac{e_p D_p \nabla_x p \mu_p^{\mathrm{H}}}{\sigma_n + \sigma_p} B - \frac{(\sigma_n \mu_n^{\mathrm{H}} + \sigma_p \mu_p^{\mathrm{H}})\, e_p D_p \nabla_x p}{(\sigma_n + \sigma_p)^2} B. \qquad (82.27)$$

In the extrinsic conductivity range the expression (82.27) is simplified since the conductivity produced by the minority carriers may be neglected. Assuming the sample to be hole-type we obtain

$$E_z = B\frac{e_pD_p\nabla_x p}{\sigma_p}\left(\mu_p^H - \frac{\sigma_p\mu_p^H}{\sigma_p}\right) = 0. \qquad (82.28)$$

Thus, *the Kikoin-Noskov field for the case when majority carriers are generated as a result of impurity absorption is zero.* Suppose now the sample is electron-type, and that minority carriers are generated in it:

$$E_z = \frac{Be_pD_p\nabla_x p}{\sigma_n}(\mu_p^H - \mu_n^H) \neq 0. \qquad (82.29)$$

It is of interest that in conditions of $E_x = 0$ the Kikoin-Noskov field is always non-zero. From (82.25) we obtain

$$E_z = \frac{e_pD_p\nabla_x p\mu_p^H B}{\sigma_n + \sigma_p} \neq 0. \qquad (82.30)$$

It follows from the expressions for E_z that the Kikoin-Noskov field in the range of small light intensities increases linearly with J. Indeed, as long as $\delta n = \delta p \ll n_0 + p_0$, $E_z \sim J$. However, as J increases, $\delta n = \delta p$ may grow to be larger than $n_0 + p_0$, and E_z will attain a saturation value the magnitude of which depends linearly on B, this linear dependence terminating for higher B's. The pattern of E_z variations in high fields may be obtained from (82.12) and (82.13) for $B \to \infty$.

The photomagnetoelectric effect is widely used in studies of various non-equilibrium processes in semiconductors. When analyzing the experimental results theoretically one should pay great attention to the experimental conditions, for different results may be obtained in different conditions.

We shall confine ourselves to one example of the practical application of the Kikoin-Noskov effect.

The distribution of the excess carrier concentration inside the sample is determined by the diffusion length

$$\delta n(x) = \delta n(0)e^{-\frac{x}{L}}. \qquad (82.31)$$

Its derivative is

$$\nabla_x n(x) = -\frac{\delta n(x)}{y} = -\frac{\delta n(0)e^{-\frac{x}{L}}}{L}. \qquad (82.32)$$

Should saturation be attained in conditions of experiment, $\delta n = \delta p \gg n_0 + p_0$, it would follow that $\sigma_n + \sigma_p = e_p\mu_p(1+b)\delta p$ and we would obtain from (82.26)

$$E_z \cong \frac{2D_n(1-b)}{(1+b)^2}\frac{1}{L} = \frac{2D_n(1-b)}{(1+b)^2\sqrt{D}}\frac{1}{\sqrt{\tau_f}}. \tag{82.33}$$

The fact that $E_z \sim \frac{1}{\sqrt{\tau_f}}$ makes *the photomagnetoelectric effect useful for measuring small lifetimes* τ_f, $\tau_f \cong (10^{-9}\text{-}10^{-10})$ s.

83. FARADAY EFFECT

Physical phenomena arising from the interaction of radiation with matter placed in magnetic fields are termed magneto-optical. They include magnetoabsorption, cyclotron resonance, plasma magnetoreflectivity, and polarization plane rotation, or the Faraday effect. Let us discuss the latter.

The existence of magneto-optical phenomena means that the optical properties of matter are affected by the magnetic field. The nature of the changes introduced by the magnetic field may in some cases be understood with the aid of Larmor's theorem. Let a quantum system (atoms, molecules, etc.) have a mechanical momentum **M** and a magnetic moment **μ** related, as usual, by the gyromagnetic factor G:

$$\boldsymbol{\mu} = G\mathbf{M}. \tag{83.1}$$

When a magnetic field **B** is applied to the system a force couple with the momentum **N** will act on the system:

$$\mathbf{N} = [\boldsymbol{\mu}\mathbf{B}], \tag{83.2}$$

which will tend to orient **μ** along the field. However, these is a mechanical momentum connected with **μ**, therefore a change in **μ** would amount to the violation of the momentum conservation law. The variation of **M** may be found by making use of the fundamental equation of rotational motion,

$$\frac{d\mathbf{M}}{dt} = \mathbf{N}, \tag{83.3}$$

or

$$\frac{d\mathbf{M}}{dt} = [\boldsymbol{\mu}\mathbf{B}] = G[\mathbf{MB}]. \tag{83.4}$$

Presuming the gyromagnetic factor to be a scalar we may write

$$d\mathbf{M} = G[\mathbf{MB}]dt, \tag{83.5}$$

i.e. the increment of the momentum $d\mathbf{M}$ is perpendicular to the momentum itself and to the magnetic field induction **B**, therefore the modulus of the vector **M** remains unchanged, and only the rotation of **M** around **B** takes place. Figure 139 shows the position of **M** in two instants of time, t and $t+dt$. The angle θ between **M** and **B** remains the same and so does the distance $\mathbf{M}_r$ between the head of the vector **M** and the **B**-axis. This means that vector **M** precesses around the direction of **B**. We are confronted with a typical gyroscopic effect.

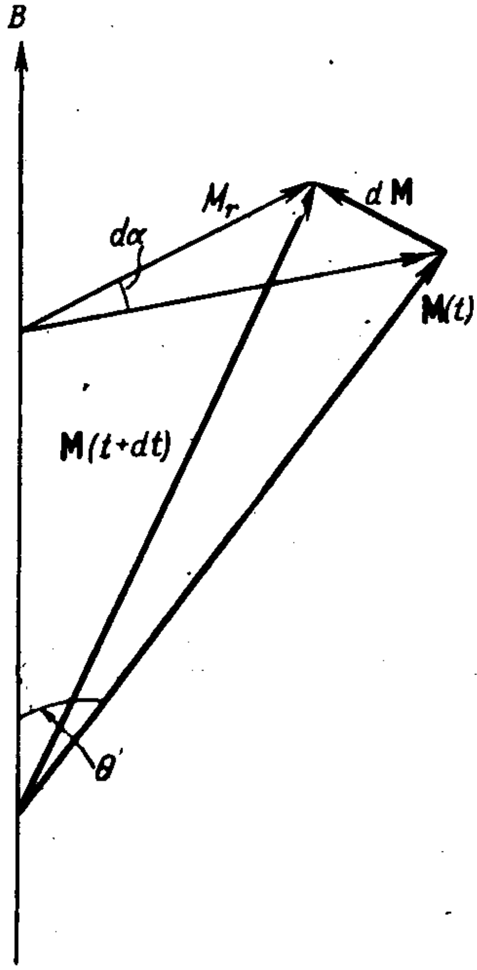


Fig. 139. The origin of momentum precession in a magnetic field

Find the angular velocity of the precession, i.e. the angular velocity of rotation of the vector $\mathbf{M}_r$ around the axis **B** (Fig. 139):

$$d\alpha=\omega_L\,dt=\frac{|d\mathbf{M}|}{|\mathbf{M}_r|}=\frac{GMB\sin\theta\,dt}{M\sin\theta}\;GB\,dt, \qquad (83.6)$$

or

$$\omega_L=GB.$$

Since these are vectorial quantities we may write the equation in vector form taking account of their orientation:

$$\omega_L=G\mathbf{B}. \qquad (83.7)$$

The angular velocity of precession ω_L is termed Larmor frequency.

In a co-ordinate system which rotates around **B** with velocity ω_L, **M** and μ will remain constant. Now we may formulate the following proposition: when a magnetic field is applied to a physical system its properties are changed in such a way that they remain constant in a rotating co-ordinate system. Apply this proposition to the Faraday effect.

It was experimentally determined that the rotation of the polarization plane of a linearly polarized light depends on the thickness l of the layer of the substance and on the magnetic field B:

$$\varphi=\theta Bl, \qquad (83.8)$$

where θ is termed Verdet coefficient. It was agreed that the rotation angle is positive if the polarization plane rotates clockwise (to the right) when the direction of light propagation coincides with the direction of the magnetic field. It follows that the rotation sign changes with the direction of the field.

The explanation of the Faraday effect is based on Fresnel's idea about the difference in the value of the refraction index for right and left circularly polarized waves. The linear polarized wave may be regarded as the superposition of two circularly polarized waves—the right (+) and the left (—) one. After passing through a layer of the thickness l the waves acquire a path difference Δ and a phase difference $\Delta\varphi$:

$$\Delta\varphi = \frac{2\pi\Delta}{\lambda} = \frac{2\pi}{\lambda}(l\mathrm{n}^+ - l\mathrm{n}^-). \qquad (83.9)$$

The angle of polarization plane rotation is equal to one half of the phase difference acquired by the waves, i.e.

$$\varphi = \frac{1}{2}\Delta\varphi = \frac{\pi}{\lambda}(\mathrm{n}^+ - \mathrm{n}^-)\, l = \frac{\omega(\mathrm{n}^+ - \mathrm{n}^-)\, l}{2c}. \qquad (83.10)$$

The rotation is possible if $\mathrm{n}^+ \neq \mathrm{n}^-$. Express n^+ and n^- in terms of n. In accordance with the above we may write

$$\mathrm{n}^\pm = \mathrm{n}(\omega \mp \omega_L). \qquad (83.11)$$

Since $\omega_L \ll \omega$, $\mathrm{n}(\omega)$ may be expanded into a series and only two of its terms retained

$$\mathrm{n}(\omega \pm \omega_L) = \mathrm{n}(\omega) \pm \frac{d\mathrm{n}}{d\omega}\omega_L \qquad (83.12)$$

and

$$\mathrm{n}^+ - \mathrm{n}^- = -2\frac{d\mathrm{n}}{d\omega}\omega_L, \qquad (83.13)$$

whence

$$\varphi = -\frac{\omega l}{2c}\frac{d\mathrm{n}}{d\omega}2\omega_L. \qquad (83.14)$$

Re-write (83.14) in the following form taking into account that $\omega\frac{d\mathrm{n}}{d\omega} = -\lambda\frac{d\mathrm{n}}{d\lambda}$:

$$\varphi = \frac{\lambda}{c}\frac{d\mathrm{n}}{d\lambda}l\omega_L = \frac{\lambda}{c}\frac{d\mathrm{n}}{d\lambda}GlB. \qquad (83.15)$$

This expression is a modification of the Becquerel relation. For the Verdet coefficient we obtain

$$\theta = \frac{\lambda}{c}\frac{d\mathrm{n}}{d\lambda}G. \qquad (83.16)$$

For the orbital motion of the electron $G = \frac{e}{2mc}$ (in the Gauss system), therefore

$$\theta = \frac{\lambda}{2c^2}\frac{e}{m}\frac{d\mathrm{n}}{d\lambda}, \qquad (83.17)$$

and

$$\varphi = \frac{\lambda}{2c^2} \frac{e}{m} \frac{dn}{d\lambda} lB. \qquad (83.18)$$

The rotation of the polarization plane takes place only if there is dispersion of light in the corresponding spectral region. But dispersion is closely related to light absorption, i.e. near the absorption line or band $\left(\frac{dn}{d\lambda} < 0\right)$ there is a region of normal dispersion, and in the absorption band itself $\left(\frac{dn}{d\lambda} > 0\right)$ the absorption is anomalous. Therefore the rotation angle should be expected to change sign as the absorption band is traversed. The sign inversion should take place twice because in traversing the absorption band $\frac{dn}{d\lambda}$ turns zero at least twice.

Since there are several types of absorption of light in semiconductors one should expect as many types of polarization plane rotation provided each absorptfon type is responsible for the refraction index dispersion.

The rotation due to dispersion caused by intrinsic absorption is termed interband rotation. It is usually observed outside the intrinsic absorption band in the region $\frac{\hbar\omega}{\Delta E_0} \sim 0.1$ to 1. The observation of interband rotation $\hbar\omega > \Delta E_0$ is hampered by intense absorption. In the same way one may consider free charge carrier and exciton rotations. The rotation by impurity states is also possible although difficult to realize in practice. The effect observed in experiment is a superposition of rotation effects caused by various mechanisms.

Consider the rotation of the polarization plane by free charge carriers. Make use of the method of the complex dielectric permeability tensor. Since the difference between n^+ and n^- is minute, we may write

$$n^+ + n^- = 2n \qquad (83.19)$$

and

$$\varphi = \frac{\pi l}{2\lambda n} (n^{+2} - n^{-2}). \qquad (83.20)$$

Substituting ε for n^2 we obtain

$$\varphi = \frac{\pi l}{2\lambda n} (\varepsilon^+ - \varepsilon^-). \qquad (83.21)$$

As was demonstrated in Sec. 74 the amplitude of the Fourier component of the radiation field satisfies the equation

$$\Delta \mathbf{E}(\mathbf{r}, \omega) + \frac{\omega^2 \mu}{e^2} \left[\varepsilon - i \frac{4\pi\sigma}{\omega}\right] \mathbf{E}(\mathbf{r}, \omega) = 0. \qquad (83.22)$$

Denote the expression in the square brackets by $\tilde{\varepsilon}$:

$$\tilde{\varepsilon}=\varepsilon-i\frac{4\pi}{\omega}\sigma, \tag{83.23}$$

i.e. introduce the complex dielectric permeability. When the magnetic field is applied, σ becomes a tensor (gyrotropy), therefore $\tilde{\varepsilon}$ is also a tensor (for the sake of simplicity we shall assume ε to be a scalar). For **B** directed along the z-axis we may write σ, according to Secs. 42, 45, in the form

$$\sigma=\begin{pmatrix}\sigma_{xx} & \sigma_{xy} & 0\\ \sigma_{yx} & \sigma_{yy} & 0\\ 0 & 0 & \sigma_{zz}\end{pmatrix}, \tag{83.24}$$

where for a weak magnetic field $\sigma_{xx}\cong\sigma_{yy}\cong\sigma_{zz}\cong\sigma(0)$; σ_{xy} is the Hall conductivity derived from the kinetic equation

$$\sigma_{xy}=-\frac{e^3}{m^*}K'_{12}B_z \tag{83.25}$$

for a finite sample. It follows from the explicit expression for σ_{xy} that $\sigma_{xy}=-\sigma_{yx}$. Indeed, the change $x\to y$ and $y\to x$, if the right-hand co-ordinate system is retained, results in $z\to -z$ and, consequently, $B_z\to -B_z$, and this proves the tensor to be antisymmetrical. Hence, we may write

$$\tilde{\varepsilon}=\begin{pmatrix}\varepsilon_{xx} & \varepsilon_{xy} & 0\\ \varepsilon_{yx} & \varepsilon_{yy} & 0\\ 0 & 0 & \varepsilon_{zz}\end{pmatrix}\quad \begin{matrix}\varepsilon_{xy}=-i\dfrac{4\pi}{\omega}\sigma_{xy}\\ \varepsilon_{yx}=-\varepsilon_{xy}\end{matrix} \tag{83.26}$$

Introducing the scalar refraction n and absorption indices nϰ with the aid of the relation

$$\mathrm{n}^2(1-i\varkappa)^2\,\mathbf{E}=\tilde{\varepsilon}\mathbf{E} \tag{83.27}$$

and equating the coefficients in front of E_x and E_y in the left-hand and right-hand sides of the equation we obtain

$$\mathrm{n}^2(1-i\varkappa)^2=\varepsilon_{xx}+\varepsilon_{yx}, \tag{83.28}$$

and

$$\mathrm{n}^2(1-i\varkappa)^2=\varepsilon_{yy}+\varepsilon_{xy}=\varepsilon_{yy}-\varepsilon_{yx}. \tag{83.29}$$

It follows from the last two equations that there are two solutions of the equations

$$\mathrm{n}^2(1-i\varkappa)^2=\varepsilon_{xx}\mp\varepsilon_{xy}, \tag{83.30}$$

since we may write for the circularly polarized wave

$$E_x=\mp iE_y. \tag{83.31}$$

(The tradition in optics is to assign the sign (+) to left- and the sign (—) to right-polarized waves. The designation of right- and left-polarized waves in quantum mechanics and in nuclear physics is exactly opposite.) Two different solutions correspond to two different polarizations

$$\mathrm{n}^{\pm 2} \cong \varepsilon_{xx} \pm \varepsilon_{xy}, \tag{83.32}$$

therefore

$$\mathrm{n}^{+2} - \mathrm{n}^{-2} = i2\varepsilon_{xy}, \tag{83.33}$$

and

$$\varphi = i\frac{\pi l}{2\mathrm{n}\lambda} 2\varepsilon_{xy} \tag{83.34}$$

or

$$\varphi = \frac{2\pi l}{\mathrm{n}c}\sigma_{xy}. \tag{83.35}$$

We may write for the Verdet coefficient

$$\theta = \frac{2\pi}{\mathrm{n}c}\frac{\sigma_{xy}}{B_z} \tag{83.36}$$

or

$$\theta = -\frac{2\pi}{\mathrm{n}c}\frac{e^3}{m^*}K'_{12}. \tag{83.37}$$

But according to (74.33)

$$\alpha = \frac{4\pi}{cn}\sigma_{xx} = \frac{4\pi}{cn}e^2K'_{11}, \tag{83.38}$$

therefore

$$\frac{\theta}{\sigma} = -\frac{2\pi e^3 K'_{12}}{ncm^* \frac{4\pi}{cn} e^2 K_{11}} = -\frac{1}{2}\frac{e}{m^*}\frac{\langle\tau^2\rangle}{\langle\tau\rangle} = -\frac{1}{2}\mu^{\mathrm{H}}, \tag{83.39}$$

i.e.

$$\theta = -\frac{\mu^{\mathrm{H}}}{2}\alpha. \tag{83.40}$$

For electrons $\mu^{\mathrm{H}} < 0$, and the Verdet coefficient is positive. For holes $\theta < 0$.

The Hall mobility μ^{H} depends on the scattering mechanism; α, too, is dependent on it, therefore θ should also depend on the scattering mechanism. Indeed, it follows directly from the expression for θ in terms of K'_{12} that

$$\theta = -\frac{2\pi}{\mathrm{n}c}\frac{e^3}{m^{*2}}p\langle\tau^2\rangle. \tag{83.41}$$

The most important features of this relation are the linear dependence of the specific rotation angle on the carrier concentration p and the inverse quadratic dependence on the effective mass m^*.

The last relation took no account of the dispersion. It is based on the phenomenological Maxwell equations and is valid in the low-frequency spectral region where $\omega^{-2} \gg \tau^2$. To obtain an expression which would hold for the high-frequency region one should take into account the factor $\frac{1}{1+\omega^2\tau^2}$ in the expressions for polarization ability (75.14) and for dielectric permeability (75.15). Inserting this factor into the expression for the Hall conductivity in weak fields we may write

$$\theta = -\frac{2\pi e^3 p}{ncm^{*2}} \left\langle \frac{\tau^2}{1+\omega^2\tau^2} \right\rangle . \tag{83.42}$$

In the high-frequency region, where $\omega^2 \ll \tau^{-2}$, we have $\omega^2\tau^2 \gg 1$; therefore, we obtain for θ

$$\theta = -\frac{2\pi e^3 p\omega^{-2}}{ncm^{*2}} = -\frac{e^3 p\lambda^2}{2\pi c^3 nm^{*2}} \tag{83.43}$$

This equation is widely used to determine the effective mass of free carriers in samples with impurity conductivity. The condition of its applicability is tested by experiment: the $\theta(\lambda^2)$, or $\varphi(\lambda^2)$, graph should be a straight line; in this case the polarization plane rotation is independent of the scattering mechanism. The deviation of the $\theta(\lambda^2)$ or $\varphi(\lambda^2)$ dependence from linear may be due either to the violation of the condition $\omega^2\tau^2 \gg 1$, or to the interference of other rotation mechanisms, the interband rotation, to take just one example.

Consider again the ratio θ/α paying attention to the frequency dependence of conductivity. According to (75.12) we may write, averaging over the energy,

$$\sigma = \frac{e^2 p}{m^*} \left\langle \frac{\tau}{1+\omega^2\tau^2} \right\rangle , \tag{83.44}$$

$$\alpha = \frac{4\pi e^2 p}{ncm^*} \left\langle \frac{\tau}{1+\omega^2\tau^2} \right\rangle , \tag{83.45}$$

and

$$\frac{\theta}{\alpha} = -\frac{e}{2m^*} \frac{\left\langle \frac{\tau^2}{1+\omega^2\tau^2} \right\rangle}{\left\langle \frac{\tau}{1+\omega^2\tau^2} \right\rangle} . \tag{83.46}$$

For $\omega^2\tau^2 \ll 1$ we obtain the already familiar expression

$$\frac{\theta}{\alpha} = -\frac{\mu^H}{2} . \tag{83.47}$$

For $\omega^2\tau^2 \gg 1$ we obtain

$$\frac{\theta}{\alpha} = -\frac{\mu^H}{2} \langle\tau\rangle^{-1} \langle\tau^{-1}\rangle^{-1} \tag{83.48}$$

In a more general case α may depend not on λ^2 but on λ^γ, where $\gamma \neq 2$, and the ratio θ/α may depend on λ.

Figure 140 shows an example of rotation by electrons in cadmium telluride. The inclination angle of the graph yielded the value $m^* = 0.09m_0$ at 100 K.

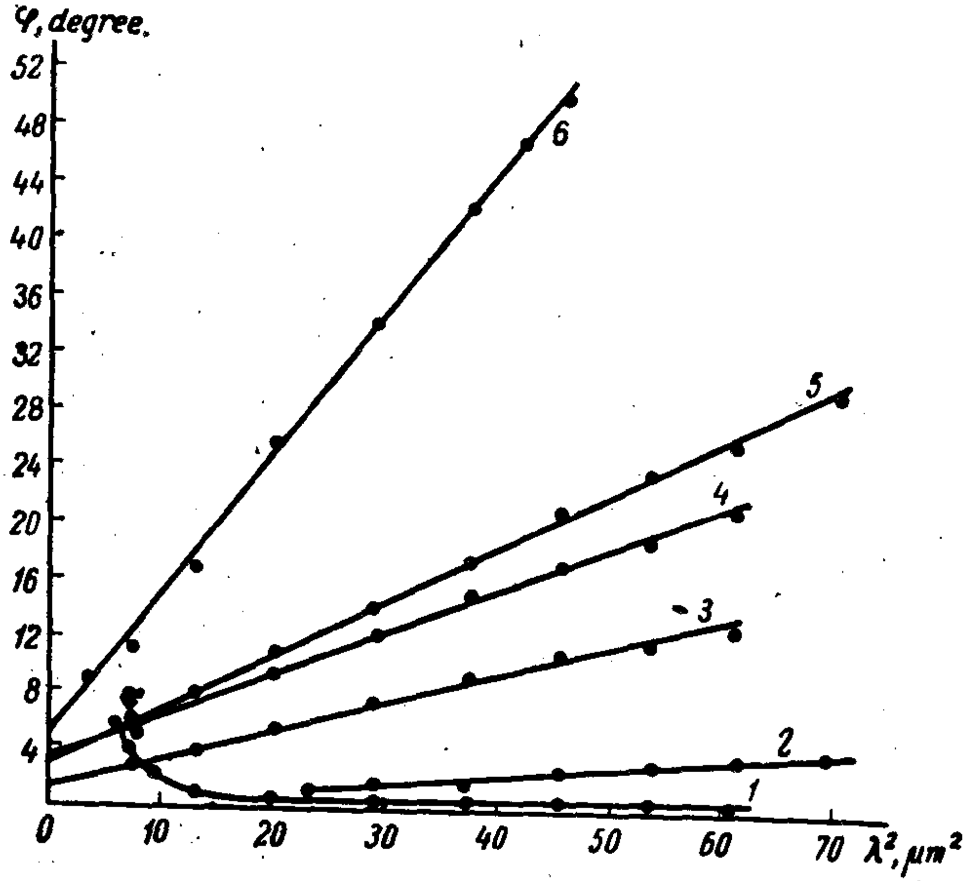


Fig. 140. Faraday effect on free electrons in cadmium telluride, ion concentration, cm:

$1 - < 10^{16}$; $2 - 2.6 \cdot 10^{16}$; $3 - 1.2 \cdot 10^{17}$; $4 - 1.5 \cdot 10^{17}$; $5 - 2.0 \cdot 10^{17}$; $6 - 3.5 \cdot 10^{17}$

The theory of interband rotation is based on the conductivity tensor. In calculating the tensor attention is paid to electron transitions from the valence band to the conduction band. The perturbation is written in the usual way:

$$\hat{W} = \sum_i e(\mathbf{v}_i \mathbf{A}') \tag{83.49}$$

where $\mathbf{v}_i$ is the velocity of the electron in the state i (in the simplest case it is equal to the ratio of the momentum or quasimomentum to the electron mass), $\mathbf{A}'$ is the vector potential of the light wave. The matrix elements of the perturbation operator are calculated with the aid of the electron wave functions for the initial **k** and the final **k**′ states of the Brillouin zone. Definite assumptions are made concerning the relations existing between the elements, for example:

$$\frac{|P^{+}_{\mathbf{kk}'}|}{\omega^{+}_{\mathbf{kk}'}}=\frac{|P^{-}_{\mathbf{kk}'}|}{\omega^{-}_{\mathbf{kk}'}}=\frac{|P_{\mathbf{kk}'}|}{\omega_{\mathbf{kk}'}}. \tag{83.50}$$

Theoretical computations supported by experiment lead to the conclusion that the frequency dependence of interband rotation is of the type ω^2 for $\omega \longrightarrow 0$. However, close to the intrinsic absorption boundary different theoretical models result in different expressions for the dependence of the rotation angle on photon energy.

According to L. Rot's computations $\varphi \sim F_1(x)$, where

$$F_1(x)=\frac{1}{x}\left(\frac{1}{(1-x)^{1/2}}-\frac{1}{(1+x)^{1/2}}\right)-\frac{4}{x^2}\left[2-(1-x)^{1/2}-(1+x)^{1/2}\right],$$
$$x=\frac{\omega}{\omega_g}=\frac{\hbar\omega}{\Delta E_0}. \tag{83.51}$$

For $x \longrightarrow 1$ the behaviour of the function $F_1(x)$ is asymptomic:

$$F_1(x)\cong\frac{1}{(1-x)^{1/2}}. \tag{83.52}$$

Near the absorption edge $\omega=\omega_g$ or $x=1$, the function $F_1(x)$ has a singularity which may be eliminated by the introduction of the relaxation time. In another theory (Bosvarva, Howard, Lydiard) the rotation is determined by the function $F_2(x)$:

$$F_2(x)=\frac{1}{x}\left[\frac{1}{(1-x)^{1/2}}-\frac{1}{(1+x)^{1/2}}\right]-1 \tag{83.53}$$

with the same asymptomic value. The behaviour of both functions for $x \longrightarrow 0$ is similar, the relation between them being $F_2(x)=2F_1(x)$. Both functions are widely used for the analysis of experimental dispersion curves usually for $x \sim (0.6$-$0.9)$. The experiment shows, however, the rotation for $x \geqslant 1$ to be finite. This result is explained in the theory of Lax *et al.* which shows the rotation φ due to dispersion caused by electron transitions between the Landau levels to be determined in case of direct transitions by the expression

$$\varphi_{\text{Landau}}=A\left\{\frac{1}{[(X-Y)^2+1]^{1/2}}\{[(X-Y)^2+1]^{1/2}-(X-Y)\}^{1/2}-\right.$$
$$\left.-\frac{1}{[(X+Y)^2+1]^{1/2}}\{[(X+Y)^2+1]^{1/2}+(X+Y)\}^{1/2}\right\}, \tag{83.54}$$

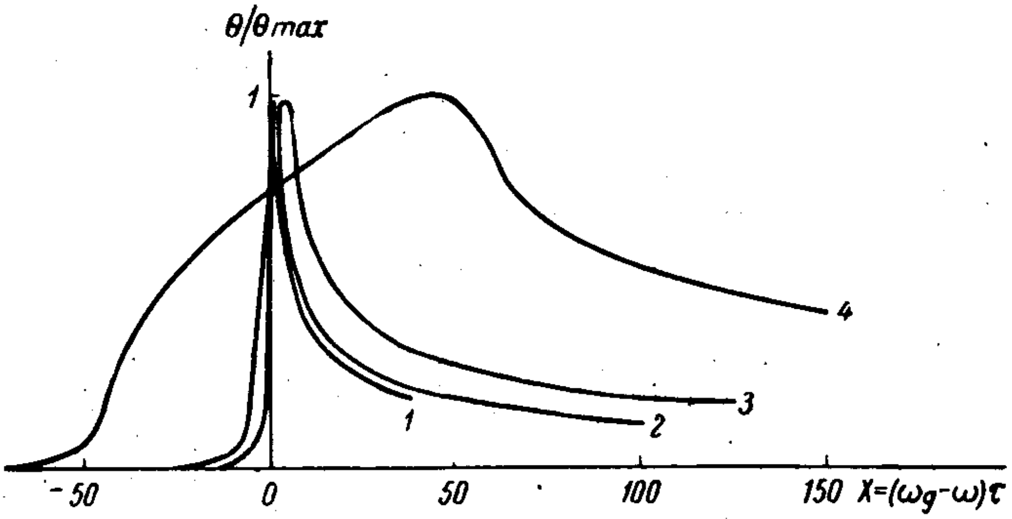


Fig. 141. A dispersion curve for direct interband transitions for different values of Y:

1—0.1; 2—1.0; 3—5.0; 4—50.0

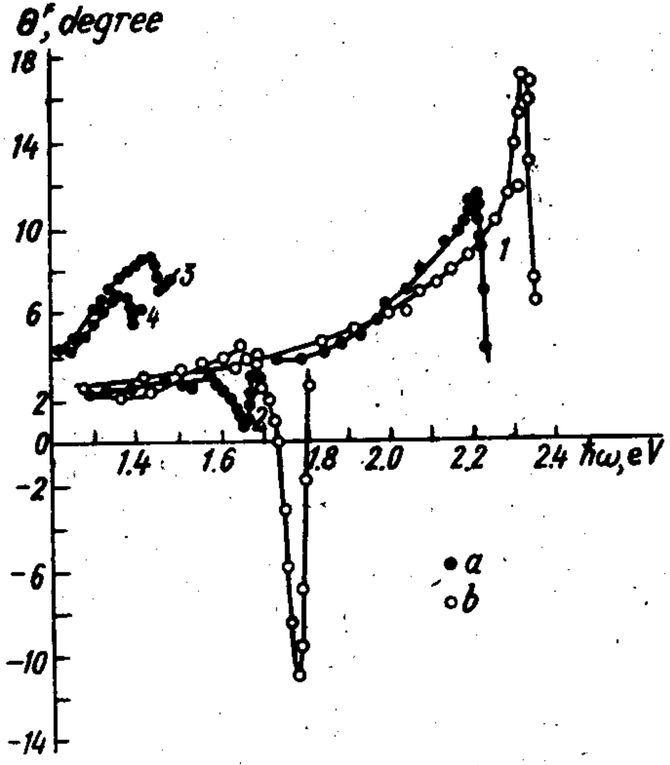


Fig. 142. Interband and exciton polarization plane rotation:

a—300 K; b—77 K; 1—ZnTe; 2—CdSe; 3—CdTe; 4—$CdTe_{0.95}Se_{0.05}$

where A is independent of frequency,

$$X=(\omega_n-\omega)\tau,\quad Y=\gamma B\tau,\quad \gamma=\frac{g_c+g_v}{2\hbar}\mu_B$$

is the effective magnetic moment expressed in terms of the gyromagnetic ratios g_c and g_v; μ_B is the Bohr magneton; $\omega_n=\omega_g+$ $+(n+1/2)\omega_c$ is the frequency of transition between the nth Landau levels, and ω_c is the cyclotron frequency $\omega_c=2\omega L$.

It may be seen from this expression that the rotation sign depends on the sign of the term $(X-Y)$—it is positive for $(X-Y)<0$, and negative for $(X-Y)>0$.

Thus, the rotation due to dispersion caused by electron transitions between two Landau levels is subject to sign inversion. The magnitude of the rotation is not symmetrical about the frequency ω_n. To take account of the rotation produced by the entire set of Landau levels one should carry out summation over n to obtain an expression of the form

$$\varphi=\frac{2A}{\omega_c\tau}\{\{[(X+Y)^2+1]^{1/2}+(X+Y)\}^{1/2}-$$
$$-\{[(X-Y)^2+1]^{1/2}+(X-Y)\}^{1/2}\},\qquad(83.55)$$

where X is now $(\omega_g-\omega)\tau$, the other notations remaining unchanged. The shape of dispersion curves for interband polarization plane rotation for some values of the parameter Y is shown in Fig. 141.

In experimental studies of the interband rotation sign inversion is observed near the absorption edge. Sometimes the inversion is observed repeatedly, the cause being attributed to exciton absorption band rotation (Fig. 142).

84. SPIN-ORBITAL SPLITTING OF ENERGY BANDS

The origin of the spin-orbital splitting of energy bands is the same as of the splitting of energy levels of isolated atoms which results in the formation of energy bands when the atoms interact with one another.

The spin-orbital splitting of atomic energy levels is usually known in atomic physics and in spectroscopy by the name of fine structure of levels (spectral lines).

The fine structure of the energy levels of free atoms and ions is well established. The theoretical background for the existence of the fine structure is presented by the Dirac relativistic quantum-mechanical equation; therefore, it may be said that the fine structure or the spin-orbital splitting is a relativistic quantum effect. The most convenient way to consider the spin-orbital interaction is to replace the Dirac equation by the Pauli equation, i.e. replace the four-component Dirac function by the two-component Pauli function. The exclusion of the "small" components from the Dirac function transforms it into an equation of the type

$$\left(\frac{\hat{\mathbf{p}}^2}{2m_0}+U(\mathbf{r})-E\right)\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}=$$

$$=\left\{\frac{\hat{\mathbf{p}}^4}{8m_0^3c^2}+\frac{e\hbar}{2m_0c}(\hat{\sigma}\mathbf{B})+\frac{e\hbar}{4m_0^2c^2}(\hat{\sigma}[\mathbf{E}\hat{\mathbf{p}}])-\frac{\hbar^2}{8m_0c^2}\nabla U\right\}\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}. \quad (84.1)$$

The right-hand side of the equation (84.1) is a relativistic quantum-mechanical correction to the non-relativistic equation in the left-hand side. The first term $\hat{\mathbf{p}}^4/8m_0^3c^2$ takes account of the relativistic dependence of the mass on the velocity. The need for this term follows from the relativistic energy-momentum relation

$$E=\sqrt{m_0^2c^4+p^2c^2}. \quad (84.2)$$

Expand $E(p)$ into a series in p^2:

$$E(p)=m_0c^2+\frac{p^2}{2m_0}-\frac{p^4}{8m_0^3c^2}+\ldots. \quad (84.3)$$

The second term is the energy of the electron in an external magnetic field $\mathbf{B}$ with which the electron interacts via its intrinsic magnetic moment μ_e:

$$\mu_e=\frac{e\hbar}{2m_0c}\hat{\sigma}=\frac{e}{m_0c}\hat{\mathbf{S}}, \quad (84.4)$$

where $\hat{\mathbf{S}}=\frac{\hbar}{2}\hat{\sigma}$ is the intrinsic mechanical moment, or spin, $\hat{\sigma}$ are 2×2 Pauli matrices.

The third term is the energy of the spin-orbital interaction. The fourth term bears the name of the contact interaction, or of the Darwin correction. Although this method of introducing spin-orbital interaction has the advantage of being rigorous it does not present a clear picture, and therefore we shall discuss the semiclassical description of such interaction.

From the relativistic point of view it is meaningless to describe electric and magnetic fields as existing independently. There is in reality one electromagnetic field described by a four-dimensional antisymmetrical field tensor $F_{\mu\nu}$:

$$\{F_{\mu\nu}\}=\begin{pmatrix}0 & B_z & -B_y & -iE_x\\ -B_z & 0 & B_x & -iE_y\\ B_y & -B_x & 0 & -iE_z\\ E_x & iE_y & iE_z & 0\end{pmatrix}. \quad (84.5)$$

The tensor components depend on the state of motion of the physical system or, to express it in a more formal way, on the co-ordinate system. If both fields—the electric and the magnetic—exist in some co-ordinate system, other co-ordinate systems may be found in which there is either no electric field or no magnetic field.

Suppose the tensor $F_{\mu\nu}$ is known in some co-ordinate system which we shall assume to be the "fixed" system (x, y, z, t), or (x, y, z, ict). We shall presume the co-ordinate system (x', y', z', t') or (x', y', z', ict') to be "moving" in the direction of the x-x' axis with a velocity $v = \beta c$. The transformation of the co-ordinate system from the "fixed" to the "moving" is performed with the aid of the Lorentz transformation

$$\begin{pmatrix} x' \\ y' \\ z' \\ ict' \end{pmatrix} = A \begin{pmatrix} x \\ y \\ z \\ ict \end{pmatrix} =$$

$$= \frac{1}{\sqrt{1-\beta^2}} \begin{pmatrix} 1 & 0 & 0 & i\beta \\ 0 & \sqrt{1-\beta^2} & 0 & 0 \\ 0 & 0 & \sqrt{1-\beta^2} & 0 \\ -i\beta & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ ict \end{pmatrix}. \tag{84.6}$$

Usual rules of tensor component transformation should be applied to find the tensor $F'_{\mu\nu}$ in the "moving" co-ordinate system:

$$F'_{\mu\nu} = \sum_{i,\, j=1}^{4} a_{\mu i} a_{\nu j} F_{ij}. \tag{84.7}$$

Suppose there is no electric field in the "fixed" co-ordinate system, $\mathbf{E} = 0$, but the magnetic field is non-zero $(\mathbf{B} \neq 0)$. The magnetic field does not act upon a stationary particle. Suppose the particle is moving with a velocity v. Introduce a co-ordinate system in which the particle would be at rest and find the field tensor in the "moving" co-ordinate system. After simple transformations of the matrix A from (84.6) we obtain for the space-like components $(\mu, \nu \neq 4)$

$$\begin{aligned} F'_{12} &= \frac{F_{12}}{\sqrt{1-\beta^2}} \quad \text{or} \quad B'_z = \frac{B_z}{\sqrt{1-\beta^2}}; \\ F'_{13} &= \frac{F_{13}}{\sqrt{1-\beta^2}} \quad \text{or} \quad B'_y = \frac{B_y}{\sqrt{1-\beta^2}}; \\ F'_{23} &= \frac{F_{23}}{\sqrt{1-\beta^2}} \quad \text{or} \quad B'_x = \frac{B_x}{\sqrt{1-\beta^2}}. \end{aligned} \tag{84.8}$$

In three-dimensional notation (84.8) may be written in the form

$$\mathbf{B}'=\frac{\mathbf{B}}{\sqrt{1-\beta^2}}. \tag{84.9}$$

For the time-like components we obtain in the same way

$$\begin{aligned} F'_{41}&=0 \quad \text{or} \quad E'_x=0;\\ F'_{42}&=\frac{-i\beta}{\sqrt{1-\beta^2}}F_{12} \quad \text{or} \quad E'_y=\frac{-\beta}{\sqrt{1-\beta^2}}B_z;\\ F'_{43}&=\frac{-i\beta}{\sqrt{1-\beta^2}}F_{13} \quad \text{or} \quad E'_z=\frac{\beta}{\sqrt{1-\beta^2}}B_y. \end{aligned} \tag{84.10}$$

In three-dimensional notation

$$\mathbf{E}'=\frac{1}{c\sqrt{1-\beta^2}}[\mathbf{vB}]. \tag{84.11}$$

It follows that a particle at rest in the "moving" co-ordinate system is acted upon by the force $e\mathbf{E}'$:

$$e\mathbf{E}'=\frac{e}{c\sqrt{1-\beta^2}}[\mathbf{vB}]. \tag{84.12}$$

For small velocities $\beta^2\ll 1$ we obtain

$$e\mathbf{E}'=\frac{e}{c}[\mathbf{vB}], \tag{84.13}$$

i.e. the expression for the Lorentz force which we have repeatedly used before.

Now consider the case when in the "fixed" co-ordinate system there is only the electric field: $\mathbf{E}\neq 0$, $\mathbf{B}=0$. Performing the computations we obtain:

$$\begin{aligned} F'_{41}&=F_{41} \quad \text{or} \quad E'_x=E_x;\\ F'_{42}&=\frac{F_{42}}{\sqrt{1-\beta^2}} \quad \text{or} \quad E'_y=\frac{E_y}{\sqrt{1-\beta^2}};\\ F'_{43}&=\frac{F_{43}}{\sqrt{1-\beta^2}} \quad \text{or} \quad E'_z=\frac{E_z}{\sqrt{1-\beta^2}}. \end{aligned} \tag{84.14}$$

For the space-like components we obtain

$$\begin{aligned} F'_{12}&=-\frac{i\beta}{\sqrt{1-\beta^2}}F_{42} \quad \text{or} \quad B'_z=\frac{\beta}{\sqrt{1-\beta^2}}E_y;\\ F'_{13}&=\frac{i\beta}{\sqrt{1-\beta^2}}F_{43} \quad \text{or} \quad B'_y=-\frac{\beta}{\sqrt{1-\beta^2}}E_z;\\ F'_{23}&=0 \quad \text{or} \quad B'_x=0, \end{aligned} \tag{84.15}$$

which may be written in three-dimensional notation as follows:

$$\mathbf{B}' = \frac{1}{c\sqrt{1-\beta^2}}[\mathbf{E}\mathbf{v}]. \tag{84.16}$$

Hence, there is a magnetic field in the "moving" co-ordinate system despite the fact that in the "fixed" co-ordinate system it was zero.

A particle with the magnetic moment μ will interact with the magnetic field $\mathbf{B}$. The energy of this interacton is equal to W:

$$W = -(\mu\mathbf{B}') = -\frac{1}{\sqrt{1-\beta^2}c}(\mu[\mathbf{E}\mathbf{v}]). \tag{84.17}$$

W is the energy of the interaction of the magnetic moment μ of a particle moving with the velocity $\mathbf{v}$ in the field $\mathbf{E}$. Express the magnetic moment μ in terms of the mechanical angular momentum S, i.e. in terms of spin:

$$\mu = g\frac{e}{2mc}\mathbf{S} = g\frac{e\hbar}{4mc}\sigma, \tag{84.18}$$

where g is the Lande factor, or the g-factor, m is the rest mass. Substituting the momentum for the velocity we obtain

$$W = -\frac{ge\hbar}{(1-\beta^2)4m^2c^2}(\sigma[\mathbf{E}\mathbf{p}]) = -\frac{ge\hbar}{4m_0^2c^2}(\sigma[\mathbf{E}\mathbf{p}]). \tag{84.19}$$

In the last transformation we made use of the fact that $m^2(1-\beta^2) = m_0^2$.

The value assigned to the g-factor linking the intrinsic magnetic moment with the mechanical angular momentum should be 2. However, setting $g=2$ we obtain an expression for W which is twice as large as the third term of the Dirac-Pauli equation (84.1). To avoid this discrepancy one should take $g=1$ (the Thomas-Frenkel correction). The Thomas-Frenkel correction means that in respect to the translational motion the intrinsic moment is analogous to the orbital moment for which $g=1$, the value of g for intrinsic moments being equal to 2. Write the expression for the interaction of the spin magnetic moment with the electric field $\mathbf{E}$ with the Thomas-Frenkel correction:

$$W = -\frac{e\hbar}{4m_0^2c^2}(\sigma[\mathbf{E}\mathbf{p}]) = \frac{\hbar}{4m_0^2c^2}(\sigma[\nabla U\mathbf{p}]). \tag{84.20}$$

Apply this expression to the motion of electrons in atoms. The intensity of the electric field of the nucleus may be expressed in the form

$$\mathbf{E} = \frac{(Z-\sigma_n)|e|}{r^3}\mathbf{r} \tag{84.21}$$

where σ_n is the screening constant, $(Z-\sigma_n)|e|$ is the charge of the nucleus screened by the inner electron shells.

Substituting (84.21) into (84.20) we obtain

$$W = -\frac{e\hbar}{4m_0^2c^2}\left(\sigma\left[\frac{(Z-\sigma_n)|e|}{r^3}\mathbf{r},\ \mathbf{p}\right]\right) =$$
$$= \frac{(Z-\sigma_n)e^2\hbar}{4m_0^2c^2r^3}(\sigma[\mathbf{rp}]) = \frac{(Z-\sigma_n)e^2\hbar^2}{4m_0^2c^2r^3}(\sigma\mathbf{L}), \qquad (84.22)$$

where $\mathbf{L}$ is the orbital mechanical angular momentum in units of $\hbar$, i.e.

$$[\mathbf{rp}] = \mathbf{M} = \hbar\mathbf{L}. \qquad (84.23)$$

For a quantum-mechanical description the operator $\hat{W}$ should be substituted for W in compliance with usual rules, and the average value of $\hat{W}$ in the state characterized by the wave function Ψ should be found. Using hydrogen-like wave functions

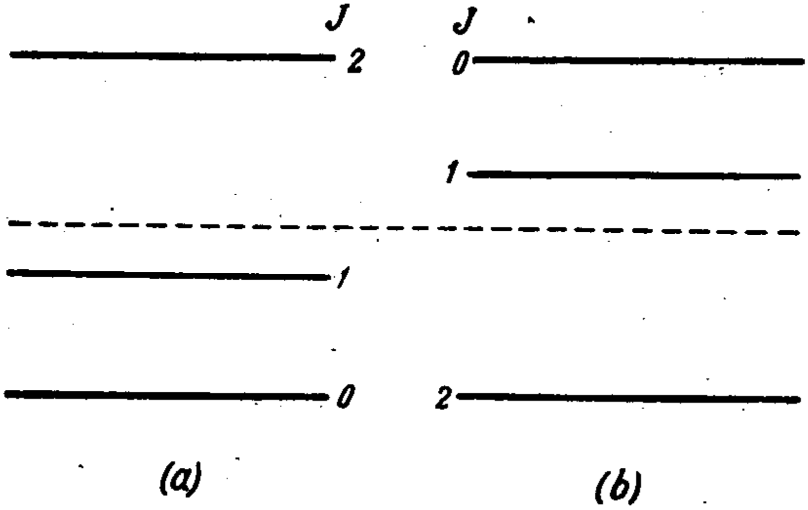


Fig. 143. Diagram of fine structure levels in atoms:
(a) normal; (b) inverse

$$\psi = \psi_{nlm}(\mathbf{r}) = R_{nl}(\mathbf{r})\,\mathbf{Y}_{lm}(\theta,\ \varphi) \qquad (84.24)$$

one may obtain

$$W = W_{nlj} = \frac{\alpha^2\,\mathrm{Ry}\,(Z-\sigma_n)^4}{n^3}\,\frac{j(j+1)-l(l+1)-s(s+1)}{2l\left(l+\frac{1}{2}\right)(l+1)}, \qquad (84.25)$$

where

$$\alpha = \frac{e^2}{\hbar c} = 7.29717\times10^{-3} \cong \frac{1}{137}$$

is the fine structure constant, Ry = 13.6 eV (rydberg), equal to Ry = Rhc, R being the Rydberg constant, n is the main, l—the orbital, and j—the full, or internal, quantum numbers.

Express W_{nlj} in electron-volts:

$$W_{nlj}\,(\mathrm{eV}) = 0.362 \times 10^{-3} \frac{(Z-\sigma_n)^4}{n^3} \frac{j(j+1)-l(l+1)-s(s+1)}{l(l+1/2)(l+1)} \tag{84.26}$$

Spin-orbital splitting should be observed for all states except the s-state since in this state the orbital angular momentum $M = \hbar L$ is zero, and according to (84.22) $W = 0$.

For hydrogen-like atoms $j = l \pm 1/2$, $s = 1/2$, therefore all the levels (except the s-level) should be doublets.

In the case of more complex atoms the quantum numbers J, L, S of the shell, as a whole, should be substituted for the one-electron quantum numbers j, l, s.

For the unexcited state of the electron shell of the group IV B and VI B atoms, $S = 1$ and $L = 1$. Therefore we have different values $J = 0$, 1 and 2 depending on the orientation of **L** and **S**. There are three different values of the orientation factor: $-4/3$, $-2/3$ and $2/3$ for $J = 0$, 1, and 2, respectively. The ratio of the distances between the fine structure levels is 1:2. Figure 143 shows a diagram of the spin-orbital splitting of the original level (dashed line) into three sublevels for the normal (*a*) and the inverse (*b*) arrangement of the levels. The orientation factor determines the number of the fine structure components, the magnitude of the splitting being determined by the shell number and by the screening constant.

Within one row of the Mendeleyev Periodic Table n and σ_n are constant, and Z increases, therefore the spin-orbital splitting increases as well. For instance, the splitting of the energy levels in atoms of the pairs C—O, Si—S, Ge—Se, Sn—Te is about the same but nevertheless it is greater for atoms of group VI B as compared to the atoms of group IV. For elements belonging to the same group n increases with the period number, Z and σ_n increasing simultaneously. However, the increase in σ_n is less rapid than in Z, therefore within one group the effective charge $(Z-\sigma_n)$ increases with Z more rapidly than n. The result is that for the elements belonging to one group the magnitude of the spin-orbital splitting increasses with the number of the element. Table 27 presents examples of spin-orbital splitting of the ground state of the atoms of some elements. For most atoms included in the table the ground state is the ${}^3P_0$ state. For selenium, however, an inversion of the levels takes place, and the ground state of tellurium is the ${}^3P_2$, the pattern of the higher levels being "normal". The interval rule is approximately valid for germanium and silicon, the splitting in tellurium being of the "double" type since the difference ${}^3P_2 - {}^3P_1$ is 0.586 eV, and ${}^3P_1 - {}^3P_0$, only 0.006 eV.

Table 27

Spin-orbital splitting of the fundamental state of the atoms of group IV and group VI elements. The deepest level is taken as the energy origin

Material		Energy of the level							
		$^3P_0$		$^3P_1$		$^3P_2$		$^3P_2$—$^3P_1$	
		cm^{-1}	eV	cm^{-1}	eV	cm^{-1}	eV	cm^{-1}	eV
Carbon	C	0.00	0.00	16.4	0.0020	43.5	0.0054	27.1	0.0033
Silicon	Si	0.00	0.00	77.15	0.009	223.16	0.0276	146.16	0.018
Germanium	Ge	0.00	0.00	557.10	0.069	1409.90	0.174	852.80	0.105
Oxygen	O	0.00	0.00	68.9	0.0084	157.5	0.0193	88.6	0.0110
Sulphur	S	0.00	0.00	176.8	0.0216	396.8	0.049	220.0	0.027
Selenium	Se	2534.4	0.317	1989.5	0.244	0.00	0.00	—544.9 ($^3P_1$—$^3P_0$)	—0.073 ($^3P_1$—$^3P_0$)
Tellurium	Te	4707	0.580	4751	0.586	0.00	0.00	—44 ($^3P_1$—$^3P_0$)	—0.006 ($^3P_1$—$^3P_0$)

As is seen from the table, the splitting of the levels of the atoms belonging to the same group actually increases with Z reaching the value of almost 0.6 eV for such comparatively heavy elements as tellurium.

The splitting of the energy bands which accompanied the formation of crystals is due to the interaction between the electron magnetic moment and the electric field of the lattice $\mathbf{E}=-\frac{1}{e}\nabla U$ The magnitude of the energy band splitting is determined by the spin-orbital splitting of the energy levels which take part in the formation of the corresponding band. Since the field of a specific atom is supplemented by the atomic interaction field, the spin-orbital splitting of the bands does not coincide with the splitting of the original energy levels. Theory and experiment show the spin-orbital splitting of the bands to be somewhat greater than that of the corresponding atomic levels. The fundamental laws governing the dependence of the spin-orbital splitting on the charge of the atomic nucleus remain the same as the charge is increased. The values of spin-orbital splitting of the energy bands for many semiconductors are at present obtained not only from experiment but from theory as well. The value of the band splitting at point Γ, i. e. for $\mathbf{k}=0$, is often denoted by Δ_0, and at point L by Δ_1, the relationship between the two in many cases being

$$\Delta_1=\frac{2}{3}\Delta_0.$$

It follows from the group theory that the spin-orbital splitting of the energy bands formed from p-levels must perforce be doublet; pure "S-bands" do not display spin-orbital splitting. This is the reason why, for example, the valence bands of the group IV

Table 28

Spin-orbital splitting of the valence band at the point Γ_{15} of some semiconductors, eV

Semiconductor	Si	Ge	ZnS	ZnSe	ZnTe	CdTe	InSb
Experimental	0.03	0.29	0.07	0.43	0.93	0.92	0.82
Theoretical	0.03	0.29	0.08	0.42	0.93	0.90	0.82

semiconductors and of the $A^{II}B^{VI}$ compounds display spin-orbital splitting and the conduction bands do not. One manifestation of this is the fact that in $A^{II}B^{VI}$ compounds the value of the splitting for a fixed B^{VI} and a variable A^{II} remains practically the same. Table 28 contains examples of spin-orbital splitting of the valence bands of some semiconductors obtained from experiment and from theory. Errors in the measurement of spin-orbital splitting do not, as a rule, exceed some hundredths fractions of an electron-volt.

Appendix

INTRODUCTION TO THE THEORY OF GROUPS

In recent years the theory of groups has been widely used in the formulation of the energy band theory and in the interpretation of experimental results. For this reason not only the theoretician but the experimenter as well should be acquainted with the concepts and notations of the group theory which are a feature of modern literature. However, for a number of reasons the author was unable to bring the whole book in line with the modern state of the art; therefore, as a stop-gap measure, it was decided to present the fundamentals of the group theory in an appendix.

1. SPACE TRANSFORMATIONS

The most important property of crystals is their symmetry, the existence of which enables several conclusions to be drawn about their physical properties with the aid of the mathematical apparatus of the group theory, because *the symmetry group* of the crystal turns out to be a group in the mathematical sense as well.

This in itself a casual fact is of the utmost importance. Before we define the group consider some space transformation operations.

Suppose we have an arbitrary point in space M with the co-ordinate $\mathbf{r}$: M $(\mathbf{r})$.

Denote a space transformation by $\hat{\mathrm{R}}$. The transformation $\hat{\mathrm{R}}$ will be understood to be a transformation which acts on the point M with the co-ordinate r and transforms its position into a new one with the co-ordinate $\mathbf{r}'$:

$$\hat{\mathrm{R}} M(\mathbf{r}) = M(\mathbf{r}'), \tag{A.1}$$

or, in short,

$$\hat{\mathrm{R}}(\mathbf{r}) = \mathbf{r}'. \tag{A.2}$$

Note, that the co-ordinate system in which the position of the point M is determined remains unchanged. There may be different transformations: $\hat{\mathrm{R}}$, $\hat{\mathrm{S}}$, $\hat{\mathrm{T}}$, ... One may introduce the concept of a product of transformations. For example, we shall term $\hat{T}$ the product of the transformations $\hat{R}$ and $\hat{S}$ if it produces the same result as the consecutive application of the transformations $\hat{R}$ and $\hat{S}$:

$$\hat{T}=\hat{S}\hat{R}. \tag{A.3}$$

Generally, the results of $\hat{R}\hat{S}$ and $\hat{S}\hat{R}$ may be different; in that case we shall have to write $\hat{R}\hat{S}\neq\hat{S}\hat{R}$. If the result is independent of the order in which the transformations are performed ($\hat{S}\hat{R}=\hat{R}\hat{S}$), the transformation is termed *commutative*.

The transformation $\hat{E}$ which does not change the space is termed *unit*, or *identical*:

$$\hat{E}M(\mathbf{r})=M(\mathbf{r}), \tag{A.4}$$

or

$$\mathbf{r}'=\hat{E}\mathbf{r}=\mathbf{r}.$$

The transformation may be multiplied by itself, for instance, $\hat{R}\hat{R}=\hat{R}^2$, or $\underbrace{\hat{R}\ldots\ldots\hat{R}=\hat{R}^p}_{p\text{ times}}$. Construct a sequence $\hat{R}$, $\hat{R}^2\ldots\hat{R}^p$. The minimum power n for which $\hat{R}^n=\hat{E}$ is termed *the order of transformation* $\hat{R}$.

Two transformations $\hat{R}$ and $\hat{R}^{-1}$ whose product is a unit transformation are termed *inverse*.

$$\hat{R}\hat{R}^{-1}=\hat{R}^{-1}\hat{R}=\hat{E}. \tag{A.5}$$

The transformation $\hat{R}^{-1}$ is termed *reciprocal to* $\hat{R}$, and $\hat{R}$ reciprocal to $\hat{R}^{-1}$:

$$(\hat{R}^{-1})^{-1}=\hat{R}. \tag{A.6}$$

The functions determined in a space may also change as the result of space transformations. For instance, if f is a function of the point M or of its co-ordinate $\mathbf{r}$,

$$f=f(M(\mathbf{r}))=f(\mathbf{r}), \tag{A.7}$$

the co-ordinate of the point and, consequently the values of the function will be changed as a result of space transformation $\hat{R}M(\mathbf{r})$. Define the action of the transformation $\hat{R}$ on the function $f(\mathbf{r})$ by the condition that in the course of the transformation the function $\hat{R}f(\mathbf{r})$ remains the same function of the point M, i. e.

$$\hat{R}f(M(\mathbf{r}))=f(\hat{R}M(\mathbf{r})), \tag{A.8}$$

or

$$\hat{R}f(\mathbf{r}) = f(\hat{R}\mathbf{r}). \qquad (A.9)$$

Hence, the transformation of space results in transformation of functions determined in this space. They may be regarded as *operators* and denoted by $\hat{R}$, $\hat{S}$, etc.

The transformations as a result of which at least one point is transformed into itself are termed *point transformations*. The appropriate point is conveniently taken as the origin of co-ordinates. Consider now some transformation types.

1. **Inversion** $\hat{i}$, i. ***A reflection at some point is termed inversion i, and the point itself—inversion centre.*** Obviously, the inversion centre is unique in that it is transformed into itself. The inversion is defined by the equation

$$\hat{i}M(\mathbf{r}) = M(-\mathbf{r}), \qquad (A.10)$$

or

$$\hat{i}\mathbf{r} = -\mathbf{r}, \ \hat{i}(x, y, z) = (-x, -y, -z). \qquad (A.11)$$

The following relation holds for the inversion operation:

$$\hat{i}f(\mathbf{r}) = f(-\mathbf{r}). \qquad (A.12)$$

It follows from the definition of inversion that it is an operation of the second order:

$$\hat{i}^2 = \hat{E}. \qquad (A.13)$$

2. **Reflection in a plane** σ. The reflection in a plane σ is an operation of the second order:

$$\hat{\sigma}^2 = \hat{E}. \qquad (A.14)$$

To formulate the reflection analytically it is necessary to determine the position of the plane. Denote a plane perpendicular to the x-axis by σ_x. The transformation $\hat{\sigma}_x$ may be represented in the form:

$$\hat{\sigma}_x\mathbf{r} = \hat{\sigma}_x(x, y, z) = (-x, y, z). \qquad (A.15)$$

By analogy

$$\begin{aligned} \hat{\sigma}_y(x, y, z) &= (x, -y, z), \\ \hat{\sigma}_z(x, y, z) &= (x, y, -z). \end{aligned} \qquad (A.16)$$

Comparing (A.15) and (A.16) with the inversion transformation we may write

$$\hat{i} = \hat{\sigma}_z\hat{\sigma}_y\hat{\sigma}_x = \hat{\sigma}_y\hat{\sigma}_z\hat{\sigma}_x = \hat{\sigma}_x\hat{\sigma}_y\hat{\sigma}_z.$$

The last two examples show that the transformation operations may be represented in a matrix form

$$\mathbf{r}' = \hat{\mathrm{R}}\mathbf{r} \tag{A.17}$$

or

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{A.18}$$

The latter relation should be accepted as the condition of equality of matrices the elements of which have been obtained in accordance with the rule of matrix multiplication; generally,

$$x_i' = \sum_j R_{ij} x_j. \tag{A.19}$$

For instance, the operations $\hat{\mathrm{I}}$ and $\hat{\sigma}$ may be represented by the matrices

$$\hat{\mathrm{I}} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}; \quad \hat{\sigma}_x = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{A.20}$$

3. Rotation axis C(φ). The rotation about an axis through the angle φ is denoted by $\hat{\mathrm{C}}(\varphi)$. The rotations through the angles $p\varphi$ may be regarded as constituting the pth power of the rotation through the angle φ:

$$\hat{\mathrm{C}}(p\varphi) = \hat{\mathrm{C}}^p(\varphi). \tag{A.21}$$

The axis $\mathrm{C}(\varphi)$ is termed n-fold axis if

$$\hat{\mathrm{C}}^n(\varphi) = \hat{\mathrm{E}}. \tag{A.22}$$

This entails the equality $n\varphi = 2\pi$. The rotation through the angle $\frac{2\pi}{n}$ is denoted by $\hat{\mathrm{C}}_n$. Obviously, the following relations hold:

$$\hat{\mathrm{C}}_n^2 = \hat{\mathrm{C}}_{n/2}; \quad \hat{\mathrm{C}}_{2n}^n = \hat{\mathrm{C}}_2; \quad \hat{\mathrm{C}}_n^{-p} = \hat{\mathrm{C}}_n^{n-p}, \text{ etc.}$$

To formulate the analytical expression of rotation the position of the rotation axis should be determined. For instance, if the rotation axis coincides with the z-axis, the co-ordinates of the points on this axis $(0, 0, z)$ are not changed in the course of the rotation through the angle φ, the z co-ordinates of the points M remain the same, and the x and y co-ordinates of all the points in space are transformed in accordance with the relation *

$$\begin{aligned} x' &= \quad x\cos p + y\sin\varphi \\ y' &= -x\sin\varphi + y\cos\varphi. \end{aligned} \tag{A.23}$$

The matrix of space rotation through the angle φ about the z-axis is written in the form

$$\hat{C}^z(\varphi) = \begin{pmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{A.24}$$

4. Rotary reflection. *The transformation of space consisting in the rotation through the angle φ and reflection in the plane $\hat{\sigma}_c$ perpendicular to the rotation axis is termed rotary reflection,* or *improper rotation,* the axis being termed *rotary reflection axis* and denoted by $\hat{S}(\varphi)$. Hence, we may write

$$\hat{S}(\varphi) = \hat{\sigma}_c\hat{C}(\varphi) = \hat{C}(\varphi)\hat{\sigma}_c. \tag{A.25}$$

The matrix of the reflection rotation about the z-axis is of the form

$$\hat{S}^z(\varphi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Even powers $2p$ of reflection rotations are identical to the pth powers of normal rotations through an angle twice the size:

$$\hat{S}^{2p}(\varphi) = [\hat{\sigma}_c\hat{C}(\varphi)]^{2p} = \hat{\sigma}^{2p}\hat{C}^{2p}(\varphi) = \hat{C}^{2p}(\varphi) = \hat{C}^{p}(2\varphi), \tag{A.26}$$

since $\hat{\sigma}^{2p} = (\hat{\sigma}^2)^p = (\hat{E})^p = \hat{E}$.

5. Inversion rotation. *The transformation of space consisting in the rotation $\hat{C}(\varphi)$ and subsequent inversion i is termed inversion rotation.*

* The transformation matrix of the co-ordinates of points of space for the rotation of the co-ordinate system through the angle φ is of the form

$$\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}.$$

This follows from the obvious fact that the rotation of the space through an angle φ is equivalent to rotation of the co-ordinate system through the angle (—φ).

Denote this transformation by $\hat{L}(\varphi)$:

$$\hat{L}(\varphi)=\hat{i}\hat{C}(\varphi)=\hat{C}(\varphi)\hat{i}. \qquad (A.27)$$

6. **Product of reflections.** A reflection in two planes σ' and σ'' inclined at an angle φ to each other results in the rotation of the space through an angle 2φ in the direction from the first plane (σ') to the second (σ'') about the intersection line of the planes:

$$\hat{\sigma}''\hat{\sigma}'=\hat{C}(2\varphi). \qquad (A.28)$$

It follows from here that

$$\hat{\sigma}'\hat{\sigma}''=-\hat{\sigma}''\hat{\sigma}'=\hat{C}(-2\varphi). \qquad (A.29)$$

7. **Product of rotations.** Suppose we have two axes C' and C'' about which rotations through the angles φ' and φ'' are performed. The product of these rotations is a rotation about a third axis C''' through the angle φ'''. The axis C''' and the angle φ''' may be found either graphically or analytically. We shall confine ourselves to an example finding $\hat{C}^y(\pi)\hat{C}^x(\pi)$.

Write

$$\hat{C}^x(\pi)=\begin{pmatrix}1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & -1\end{pmatrix},\quad \hat{C}^y(\pi)=\begin{pmatrix}-1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & -1\end{pmatrix}. \qquad (A.30)$$

Then

$$\hat{C}^y(\pi)\cdot\hat{C}^x(\pi)=\begin{pmatrix}-1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & -1\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & -1\end{pmatrix}=$$
$$=\begin{pmatrix}-1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & 1\end{pmatrix}=\hat{C}^z(\pi). \qquad (A.31)$$

A plane orthogonal to an axis is termed horizontal and often denoted by σ_h. A plane which contains the rotation axis is termed vertical and denoted by σ_v.

Note that point transformations may be expressed one in terms of the other.

8. **Translation.** The point transformations are a special case of space transformations. Another special case is *translation. It is defined as the displacement of space by an arbitrary vector* **n**. Denoting it by $\hat{T}(\mathbf{n})$ we write

$$\hat{T}(\mathbf{n})\mathbf{r}=\mathbf{r}'=\mathbf{r}+\mathbf{n}. \qquad (A.32)$$

The displacement of space by the vector **n** results in the displacement of the radius vector of the point $M(\mathbf{r})$ by the same value. Obviously

$$\hat{T}(\mathbf{n}) f(\mathbf{r}) = f(\mathbf{r}+\mathbf{n}) \qquad (A.33)$$

and

$$\hat{f}(\mathrm{n})\hat{T}(\mathrm{n}) = \hat{T}^2(\mathrm{n}) = \hat{T}(2\mathrm{n}). \qquad (A.34)$$

Moreover,

$$\hat{T}(\mathrm{n}')\hat{T}(\mathrm{n}'') = \hat{T}(\mathrm{n}'+\mathrm{n}'') = \hat{T}(\mathrm{n}'')\hat{T}(\mathrm{n}'). \qquad (A.35)$$

9. Rotation and translation. *A transformation which involves the rotation of the space through the angle* φ *about some axis and the subsequent translation by the vector* **n** *is termed rotation-translation transformation and denoted by* $\{\hat{\varphi}/\mathbf{n}\}$.

This means that

$$\{\hat{\varphi}/\mathbf{n}\}\,\mathbf{r} = \mathbf{r}' = \hat{\varphi}\mathbf{r} + \mathbf{n} = \hat{C}(\varphi)\,\mathbf{r} + \mathbf{n}, \qquad (A.36)$$

where $\hat{\varphi}$ is the matrix that corresponds to the rotation $\hat{C}(\varphi)$. It is essential that the rotation should be carried out first, followed by the translation. A special case of the rotation-translation transformation is the translation along the rotation axis. In this case the rotation axis is termed *screw axis*.

Reflection $\hat{\sigma}$ followed by translation **n** may also be considered: $\{\hat{\sigma}/\mathbf{n}\}$.

If the plane contains the translation vector **n** it is termed *glide plane*.

Let us confine ourselves to the examples of combinations of various point transformations with translation and find the product of two space transformations $\{\hat{\varphi}'/\mathbf{n}'\}$ and $\{\hat{\varphi}''/\mathbf{n}''\}$:

$$\begin{aligned}\{\hat{\varphi}''/\mathbf{n}''\}\{\hat{\varphi}'/\mathbf{n}'\}\,\mathbf{r} &= \{\hat{\varphi}''/\mathbf{n}''\}(\hat{\varphi}'\mathbf{r}+\mathbf{n}') = \\ &= \hat{\varphi}''(\hat{\varphi}'\mathbf{r}+\mathbf{n}) + \mathbf{n}'' = \hat{\varphi}''\hat{\varphi}'\mathbf{r} + \hat{\varphi}''\mathbf{n}' + \mathbf{n}'',\end{aligned} \qquad (A.37)$$

or

$$\{\hat{\varphi}''/\mathbf{n}''\}\{\hat{\varphi}'/\mathbf{n}'\} = \{\hat{\varphi}''\hat{\varphi}'/\hat{\varphi}''\mathbf{n}'+\mathbf{n}''\}. \qquad (A.38)$$

2. GROUP OF SYMMETRY TRANSFORMATIONS. PROPERTIES OF GROUP ELEMENTS

In the preceding section space transformations were considered which were subject to no limitations. Now we shall consider transformations in a crystal space by which we will understand a space containing a specified system of regular points constituting a crystal lattice.

Transformations of symmetry is the term applied to such transformations which make a space coincide with itself. Since all the equivalent lattice sites are occupied by identical atoms or groups of atoms which are impossible, in principle, to be distinguished from another such atom or group, the transformed crystal space will be indistinguishable from the original. This enables a more general definition of the symmetry transformations to be made: *the symmetry transformations of the crystal space are such transformations which leave it invariant*. The totality of symmetry transformations of the crystal space constitutes its *symmetry group*.

The following features are characteristic of this group:

1. The totality includes the unit transformation $\hat{E}$.

2. There is a reciprocal transformation $\hat{R}^{-1}$ for each transformation $\hat{R}$ whose consecutive application is equivalent to the unit transformation.

3. The consecutive application of two transformations $\hat{R}$ and $\hat{S}$ yields the same result as the application of the transformation $\hat{T}$ belonging to the same totality. In this sense the transformation $\hat{T}$ may be termed product of the transformations: $\hat{S}\hat{R}=\hat{T}$. At the same time this means that the transformation totality is a complete one.

The properties enumerated above correspond to the definition of an abstract group in mathematics (by group axioms or postulates).

Group is the term applied to a set of G elements R S, T, ... which satisfy the following conditions:

1. There is a unique element G_m belonging to the same set G as a pair of elements G_l and G_k and corresponding to them. The rule governing this correspondence is called the *composition law*, or the definition of the product of elements: $G_m=G_lG_k$.

2. The set G includes such an element E which satisfies the condition

$$EG_l=G_lE=G_l. \qquad (A.39)$$

The element E is termed *unit*, or *identical*.

3. There is a reciprocal element G_l^{-1} of each element G_l belonging to the group G for which

$$G_lG_l^{-1}=G_l^{-1}G_l=E. \qquad (A.40)$$

4. The product of the elements is associative:

$$G_lG_kG_l=G_l(G_kG_l)=(G_lG_k)G_l. \qquad (A.41)$$

The latter condition means that the result of the multiplication of any number of elements remains the same if any pair of neighbouring elements is replaced by their product.

Should we consider the set of symmetry transformations of a crystal space, we would arrive at the conclusion that it satisfies all the group postulates, and for this reason the symmetry transformation group may be regarded as an example of a mathematical group.

Here are some definitions:

1. The elements of the group G are termed *commutative* if they satisfy the relation

$$G_i G_k = G_k G_i. \tag{A.42}$$

2. A group every pair of whose elements commutes is termed *Abelian*.

3. A group is termed *finite* if it contains a finite number of elements g. The number of elements g in a group is termed *order of the group*.

4. A set of elements G_i, G_i^2, G_i^3 ... $G_i^n = E$ is *the cycle of the element* G_i, n being *the order of the element* G_i.

5. If every element belonging to the group G may be represented in the form of some integral power of any element, the group is termed *cyclic*. A cyclic group is, of necessity, an Abelian group:

$$G_i^{l_1} \cdot G_i^{l_2} = G_i^{l_1 + l_2} = G_i^{l_2} \cdot G_i^{l_1}. \tag{A.43}$$

6. A part H of the group G elements which satisfies all the group postulates is termed *subgroup*. The order of the subgroup h is a divisor of the group order g:

$$\frac{g}{h} = l, \tag{A.44}$$

where l is an integer (*Lagrange theorem*).

7. The cycle of each element G_i forms a subgroup of the group G, and for this reason the order of the element G_i is a divisor of the group order g. If the group order is a prime number such a group is obligatory cyclic.

8. The reciprocal element of a product RST... is made up of the product of corresponding reciprocal elements applied in an inverted order:

$$(RST\ldots)^{-1} = (\ldots T^{-1} S^{-1} R^{-1}). \tag{A.45}$$

Direct multiplication of the original element and the reciprocal element may be applied to prove this proposition.

Examples of the properties described have been actually given above. Here we would only like to point out that a group of translations and a group of point symmetry transformations may be regarded as subgroups of a space group.

9. A concept of paramount importance for the application of the group theory is the concept of a *class*, or of a *class of conjugate elements*.

This concept is introduced as follows. Take a definite element $T \in G$ and some arbitrary element G_i and construct a product of these elements in the form

$$G_iTG_i^{-1}.$$

This is the *similitude, or conjugate, product*. Naturally, $G_iTG_i^{-1}$ will, too, be an element of the group. Suppose it is the element S. *The element S is termed conjugate to the element T by the element G_i.* Obviously, the element T, in turn, will be conjugate to the element S by the element G_i^{-1}. Indeed, premultiply the initial relation $S=G_iTG_i^{-1}$ by G_i^{-1} and postmultiply it by G_i to obtain

$$G_i^{-1}SG_i=(G_i^{-1})S(G_i^{-1})^{-1}=G_i^{-1}G_iTG_i^{-1}G_i=ETE=T. \quad \text{(A.46)}$$

Let now the element G_i take all values of the group G. We shall obtain g products of the form $G_iTG_i^{-1}$. Some of them may be identical. For instance, the elements of the form ETE^{-1} and TTT^{-1} coincide with T. If G_k commutes with T it, too, will not produce a new element:

$$G_kTG_k^{-1}=G_kG_k^{-1}T=ET=T. \quad \text{(A.47)}$$

Therefore, the number of different elements not coinciding with T and belonging to the set $G_iTG_i^{-1}$; $G_i \in G$; $i=1, 2, \ldots, g$, will certainly be less than g. Taking only the different elements we obtain a set of elements conjugate to T. It is termed *class of elements conjugate to T*, or *class of element T*. The elements of the class T are all mutually conjugate.

Indeed, suppose the elements R and S are conjugate to the element T by the elements U and V^{-1}:

$$R=UTU^{-1}, \quad S=V^{-1}T(V^{-1})^{-1}=V^{-1}TV.$$

Expressing T in terms of S

$$T=VSV^{-1}$$

and substituting VSV^{-1} for T in R we obtain

$$R=UVSV^{-1}U^{-1}=UVS(UV)^{-1}. \quad \text{(A.48)}$$

We made use of the fact that $(UV)^{-1}=V^{-1}U^{-1}$. The number of elements in a class is termed *order of the class*. All the elements of the group may be uniquely distributed over the classes in a unique way. Every element may belong only to one class. The number of classes in a group is a very important characteristic of the group.

10. In the Abelian groups the number of classes is equal to the group order. Indeed, from the commutation rule for all the elements of the group

$$G_iG_k = G_kG_i$$

it follows that

$$G_iG_kG_i^{-1} = G_k,$$

i.e. that every element G_k is conjugate only to itself, constitutes its own class, and for this reason the number of classes is equal to the number g of elements in the group G. For the same reason the unit element constitutes its own class:

$$G_iEG_i^{-1} = EG_iG_i^{-1} = EE = E.$$

It also follows that no other class may include the unit element, therefore no class may be a subgroup.

The elements of the same class are of the same order. Indeed, let $R = UTU^{-1}$ and $T^n = E$. Consider

$$R^p = \underbrace{(UTU^{-1})(UTU^{-1})\ldots(UTU^{-1})}_{p} = UT^pU^{-1}. \qquad \text{(A.49)}$$

For $p = n$ $T^n = E$) and $R^n = UEU^{-1} = E$; for $p < n$ $T^p \neq E$ and $R^p \neq E$.

Elements of the same order may, however, belong to different classes.

3. RELATION BETWEEN GROUPS

The primary reason for which the mathematical theory of groups is of such importance is that it is applicable to groups of unspecified nature since it treats their most general properties of the group elements. One may construct the probable properties of the groups on the basis of the properties of the elements of an abstract group, and afterwards apply them to 'specific groups with the aid of the group isomorphism concept. *Two groups $G^{(1)}$ and $G^{(2)}$ of the same order are called isomorphic if each element of one group may be made to correspond to one and only one element of the second so that if $G_i^{(1)} \leftrightarrow G_i^{(2)}$ and $G_k^{(1)} \leftrightarrow G_k^{(2)}$, then $G_i^{(1)}G_k^{(1)} \leftrightarrow G_i^{(2)}G_k^{(2)}$.* All isomorphic groups are identical in everything except the nature of their elements.

Consider an example. A second-order group may be expressed in the form $G: E$, A, where $A^2 = E$, $AE = A$. An abstract group (E, A) may be realized in the form of three specific groups: $C_i: E$, i (the inversion group); $C_s: E$, σ (the reflection group), and $C_2: E$, C_2 (the rotation through the angle π group). The element E and the elements C_2, i, σ of each group correspond to each other.

A third-order group, $G{:}E, A, B$ must, of necessity, be cyclic. It may be realized as a group of rotations about a three-fold axis C_3 with the elements E, C_3 and C_3^2. The elements C_3 and C_3^2 are inverse:

$$C_3^{-1} = C_3^3 \cdot C_3^{-1} = C_3^2; \quad C_3^{-2} = C_3^4 = C_3. \qquad \text{(A.50)}$$

A fourth-order group G: E, A, B, C may belong to one of two types. Firstly, it may be of the cyclic type; for instance, $B = A^2$, $C = A^3$, $A^4 = E$. In this case it contains 4 classes: E, A, B, C. Note, that it may be regarded as a cycle of any elements A, B or C, each of them being a fourth-order element. In addition to a cyclic group, a group may be constructed of second-order elements: E, A, B, C; $A^2 = E$, $B^2 = E$, $C^2 = E$. Find the relationship between the elements A, B, C.

Consider the product AB. This is a second-order element, since $A^2B^2 = EE = E$.

$$(AB)^{-1} = A^{-1}B^{-1} = A^{-1}B^{-1} \cdot A^2B^2 = AB$$
$$(BA)^{-1} = B^{-1}A^{-1} = B^{-1}A^{-1} \cdot B^2A^2 = BA \qquad \text{(A.51)}$$

Hence, $AB = BA$. AB cannot coincide with A or B, neither can it coincide with E, therefore only one possibility remains: $AB = C$, whence we obtain $A = CB = BC$ and $B = AC = CA$. Two fourth-order groups one of which consists of fourth-order elements and the other of second-order elements are not isomorphic.

The properties of an abstract group may conveniently be illustrated by a multiplication table. It is a table of all possible products of the group elements, the first multiplier of which should be taken from the top line, and the second, from the left column. For a cyclic group of fourth-order elements it assumes the form

$G^4_{(1)}$	E	A	B	C
E	E	A	B	C
A	A	B	C	E
B	B	C	E	A
C	C	E	A	B

(A.52)

For a group of second-order elements

$G^{(4)}_{(2)}$	E	A	B	C
E	E	A	B	C
A	A	E	C	B
B	B	C	E	A
C	C	B	A	E

(A.53)

The tables show the groups to be, indeed, non-isomorphic. Any fourth-order group should be isomorphic with one of the former groups. For instance, the groups C_4 and S_4, i.e. the groups of fourth-order rotations and rotary-reflections are isomorphous with the cyclic group $G^{(4)}_{(1)}$. Let us try to construct a group of second-order elements; i, σ or C_2 may be chosen for such elements. Choose first two elements i and σ. The third element of the second order will be $i\sigma$. Hence, the group will be of the form: E, i, σ, $i\sigma$. Consider the element $i\sigma$. Since i may be represented in the form of three reflections in the co-ordinate planes, $i=\sigma_x\sigma_y\sigma_z$, we choose the given plane σ for the plane σ_z: $\sigma=\sigma_z$. In this case we obtain

$$i\sigma = i\sigma_z = \sigma_x\sigma_y\sigma_z\sigma_z = \sigma_x\sigma_y. \qquad \text{(A.54)}$$

But $\sigma_x\sigma_y$ is a rotation about the intersection line of these planes (in this case it is the z-axis) through an angle equal to twice the angle between the planes. Since in this case the angle is $\frac{\pi}{2}$, the line will constitute a two-fold axis C_2. Generally, it may be said that $i\sigma=C_2$, where C_2 is an axis orthogonal to the plane σ. Hence, we may write the group in the form $A=i$, $B=\sigma$, $C=C_2=i\sigma$. Taking account of the isomorphism we may write the multiplication table in the form

$G^{(4)}_{(2)}$	E	i	σ	C_2
E	E	i	σ	C_2
i	i	E	C_2	σ
σ	σ	C_2	E	i
C_2	C_2	σ	i	E

(A.55)

This leads us to an important conclusion: if a group contains a C_2-axis and i, it should also have a reflection plane orthogonal to C_2, and vice versa, the existence of a reflection plane and of a C_2-axis implies the existence of inversion i.

Find the classes of the group. Write the elements of the i class in the following form: i, $\sigma i\sigma^{-1}$ and $C_2iC_2^{-1}$. Resorting to the multiplication table we obtain

$$\sigma i\sigma^{-1} = \sigma i\sigma = C_2\sigma = i; \qquad \text{(A.56)}$$

$$C_2iC_2^{-1} = C_2iC_2 = \sigma C_2 = i. \qquad \text{(A.57)}$$

Hence, the class of the i element consists only of i itself. In the same way we obtain that C_2, as well as σ, makes a class in itself.

A fifth-order group must, of necessity, be cyclic: $A^5=E$. Any cyclic group of the nth order is isomorphous with a group in which

$$A=\sqrt[n]{1}=e^{i\frac{2\pi}{n}}.$$

In addition to the one-to-one correspondence of the elements belonging to two groups there may be a correspondence in which one element of the group corresponds to several of the other. In this case the groups are termed homomorphic.

A new group G may be formed from the two groups $G^{(1)}$ and $G^{(2)}$ in the following way. Construct $g=g_1g_2$ binary products of the elements $G_i^{(1)}G_k^{(2)}=G_k^{(2)}G_i^{(1)}$. The set made up of g such binary products satisfies all the group postulates, and for this reason G is a group. It is termed direct product of the groups $G^{(1)}$ and $G^{(2)}$ and denoted by

$$G=G^{(1)}G^{(2)}.$$

The order of the group $G=G^{(1)}G^{(2)}$ is equal to the product of the orders of the groups multipliers: $g=g_1g_2$. The number of classes r is equal to the product of the numbers of classes of the respective groups: $r=r_1r_2$. Consider some examples. Find the direct product of the groups C_s and C_i

$$\begin{gathered} C_s\colon E,\ \sigma;\ C_i\colon E,\ i \\ C_sC_i\colon E,\ \sigma,\ i,\ \sigma i. \end{gathered} \tag{A.58}$$

This fourth-order group $C_{(2)}^{(4)}$ has already been constructed above.

4. REPRESENTATION OF GROUPS

Most important in quantum mechanics and in solid-state physics is the *theory of group representation.*

Representation of an abstract group G is the term applied to a group of square matrices Γ homomorphic with the group G. It follows from the definition of homomorphism that there is a matrix $\Gamma(G_i)$ belonging to the group Γ to correspond to each element G_i of the group G, so that if the matrices $\Gamma(G_i)$ and $\Gamma(G_k)$ correspond to the elements G_i and G_k, the matrix corresponding to the element G_iG_k will be $\Gamma(G_iG_k)$. This condition enables us to write the main property of the representation matrices:

$$\Gamma(G_iG_k)=\Gamma(G_i)\,\Gamma(G_k). \tag{A.59}$$

If A is a square matrix of the same dimensionality f as the matrix $\Gamma(G_i)$, one may obtain the matrices $\Gamma'(G_i)$ by applying the similitude transformation to the matrices $\Gamma(G_i)$:

$$\Gamma'(G_i)=A\Gamma(G_i)\,A^{-1}. \tag{A.60}$$

The matrices $\Gamma'(G_i)$ are a representation of the group G just like the matrices $\Gamma(G_i)$. The representations $\Gamma(G_i)$ and $\Gamma'(G_i)$ are termed *equivalent*.

It follows from the definition of the representation that the reciprocal element is represented by the reciprocal matrix. Indeed, it follows from the identity

$$EG_i = G_iE = G_i$$

that

$$\Gamma(EG_i) = \Gamma(E)\,\Gamma(G_i) = \Gamma(G_i)\,\Gamma(E) = \Gamma(G_i) \tag{A.61}$$

and that, consequently, $\Gamma(E)$ is a unit matrix, i.e. that the unit element is represented by a unit matrix. In this case it follows from $E = G_iG_i^{-1}$ that

$$\Gamma(E) = \Gamma(G_i^{-1}G_i) = \Gamma(G_i^{-1})\,\Gamma(G_i) = \Gamma(G_i)\,\Gamma(G_i^{-1}),$$

i.e. that

$$\Gamma(G_i^{-1}) = \Gamma^{-1}(G_i). \tag{A.62}$$

It is easiest of all to find a reciprocal matrix for a unitary matrix. Recall that a matrix U is termed unitary if the reciprocal matrix U^{-1} is equal to the conjugate U^+,.where $U^+ = \tilde{U}^*$, the indices $\sim$ and * denoting transposition and complex conjugation of the matrix U. One may write for a unitary matrix:

$$\begin{aligned} U^{-1} &= U^+ \\ (U^{-1})_{ik} &= U^*_{ki}, \end{aligned} \tag{A.63}$$

Moreover, it follows from

$$E = U^{-1}U = UU^{-1} = U^+U = UU^+ \tag{A.64}$$

that

$$\begin{aligned} \sum_k \tilde{U}^*_{ik}U_{kj} &= \sum_k U^*_{ki}U_{kj} = \delta_{ij}; \\ \sum_k \tilde{U}_{ik}U^*_{ki} &= \sum_k U_{ik}U^*_{jk} = \delta_{ij}, \end{aligned} \tag{A.65}$$

i.e. the lines and columns of a unitary matrix may be regarded as orthonormalized vectors in f-dimensional space.

Unitary representations which place no limitations on generality are in common use:

$$\Gamma(G_i^{-1}) = \Gamma^{-1}(G_i) = \Gamma^+(G_i) = \tilde{\Gamma}^*(G_i). \tag{A.66}$$

A number of new representations may be obtained from one representation or from a combination of several representations. One of the methods of combining new representations is the construction of supermatrices. Let $\Gamma^{(1)}(G_i)$, $\Gamma^{(2)}(G_i)$, ..., $\Gamma^{(l)}(G_i)$ be some representations of the element G_i of the group G. Construct a supermatrix

$$\Gamma(G_i)=[\Gamma^{(1)}(G_i),\ \Gamma^{(2)}(G_i),\ \ldots,\ \Gamma^{(l)}(G_i)]=$$
$$=\begin{pmatrix} \Gamma^{(1)}(G_i) & 0 & 0 & 0 \\ 0 & \Gamma^{(2)}(G_i) & 0 & 0 \\ 0 & 0 & \Gamma^{(3)}(G_i) & 0 \end{pmatrix}. \qquad (A.67)$$

Such a matrix is also termed quasidiagonal. The product of two supermatrices of the same structure is also a supermatrix of the same structure. It may easily be checked with the aid of the usual matrix multiplication rule, that

$$\begin{pmatrix} \Gamma^{(1)}(G_i) & 0 & 0 & 0 \\ 0 & \Gamma^{(2)}(G_i) & 0 & 0 \\ 0 & 0 & \Gamma^{(3)}(G_i) & 0 \\ 0 & 0 & 0 & \Gamma^{(4)}(G_i) \end{pmatrix}\times$$
$$\times\begin{pmatrix} \Gamma^{(1)}(G_k) & 0 & 0 & 0 \\ 0 & \Gamma^{(2)}(G_k) & 0 & 0 \\ 0 & 0 & \Gamma^{(3)}(G_k) & 0 \\ 0 & 0 & 0 & \Gamma^{(4)}(G_k) \end{pmatrix}=$$
$$=\begin{pmatrix} \Gamma^{(1)}(G_iG_k) & 0 & 0 & 0 \\ 0 & \Gamma^{(2)}(G_iG_k) & 0 & 0 \\ 0 & 0 & \Gamma^{(3)}(G_iG_k) & 0 \\ 0 & 0 & 0 & \Gamma^{(4)}(G_iG_k) \end{pmatrix} \qquad (A.68)$$

This means that a supermatrix combined from representation matrices is a representation of a group. Moreover, this fact forms a basis for the supermatrix multiplication rule:

$$\Gamma(G_i)\Gamma(G_k)=[\Gamma^{(1)}(G_i);\ \Gamma^{(2)}(G_i);\ \ldots;\ \Gamma^{l}(G_i)]\times$$
$$\times[\Gamma^{(1)}(G_k);\ \Gamma^{(2)}(G_k);\ \ldots;\ \Gamma^{(l)}(G_k)]=$$
$$=[\Gamma^{(1)}(G_i)\Gamma^{(1)}(G_k);\ \Gamma^{(2)}(G_i)\Gamma^{(2)}(G_k);\ \ldots;\ \Gamma^{(l)}(G_i)\Gamma^{(l)}(G_k)]=$$
$$=[\Gamma^{(1)}(G_iG_k);\ \Gamma^{(2)}(G_iG_k);\ \ldots;\ \Gamma^{(l)}(G_iG_k)]=\Gamma(G_iG_k). \qquad (A.69)$$

Now apply to the supermatrices the similitude transformation with the aid of an arbitrary matrix A of the same dimensionality f as the supermatrices. We obtain some equivalent representation $\Gamma_i'=A\Gamma A^{-1}$ which may be of a non-quasidiagonal form. For this reason the appearance of the matrices $\Gamma'(G_i)$ will not warrant the statement that the representation $\Gamma'(G_i)$ has been obtained from other representations. However, it is possible with the aid of a similitude transformation by the matrix A^{-1} to obtain from $\Gamma'(G_i)$

a matrix $\Gamma(G_i) = A^{-1}\Gamma'(G_i)A$ which will be of a quasidiagonal form.

A representation which may be reduced to the quasidiagonal form with the aid of some matrix A called reducing matrix is termed reducible. If, on the other hand, there is no such matrix A which would reduce the representation to the quasidiagonal form, it is termed irreducible. The number of reducible representations is unlimited, the number of irreducible representations, on the other hand, is limited.

5. THE PROPERTIES OF IRREDUCIBLE REPRESENTATIONS

Below we offer without proof several statements pertaining to irreducible representations.

1. The number of irreducible representations of a group G is equal to the number of its classes r. It was already noted that the number of classes in a group is its most important characteristic, and the latter fact was the reason for it.

2. The relationship between the dimensionalities f_α of the irreducible representations $\Gamma^{(\alpha)}(G_i)$ and the order of the group g is as follows:

$$\sum_{\alpha=1}^{r} f_\alpha^2 = g. \tag{A.70}$$

3. The matrix elements $\Gamma^{(\alpha)}_{\mu\nu}(G_i)$ of irreducible representations constitute a system of orthonormalized vectors in the space of the group G elements, and this may be written in the form

$$\sum_{G_i} \Gamma^{(\alpha)*}_{\mu\nu}(G_i)\,\Gamma^{(\beta)}_{\mu'\nu'}(G_i) = \frac{g}{f_\alpha}\delta_{\alpha\beta}\delta_{\mu\mu'}\delta_{\nu\nu'}. \tag{A.71}$$

This is the fundamental property of the irreducible representation matrix from which all the other properties may be derived. The equation written above reflects orthogonality in various representations (α and β), lines (μ, μ') and columns (ν, ν'). To normalize the equation both sides of it should be multiplied by $\frac{f_\alpha}{g}$.

Using the similitude transformation one may vary the form of the irreducible representation matrices. However, despite the fact that the matrix elements $\Gamma^{(\alpha)}_{\mu\nu}(G_i)$ are changed in the process, the orthogonality relations remain valid. This fact is reflected in the existence of an invariant quantity termed character of the element.

The character of the group element in a specific representation is the sum of the matrix elements of the representation:

$$X(G_i) = \sum_{\mu=1}^{f} \Gamma_{\mu\mu}(G_i), \tag{A.72}$$

i.e. the element character is the trace of the matrix $\mathrm{Sp}\,\Gamma(G_i)$, or $\mathrm{Tr}\,\Gamma(G_i)$.

Check whether this quantity remains unchanged in the similitude transformation:

$$\Gamma'(G_i) = A\Gamma(G_i)A^{-1}; \tag{A.73}$$

$$X'(G_i) = \sum_{\mu=1}^{f} \Gamma'_{\mu\mu}(G_i) = \sum_{\mu,j,k} A_{\mu j}\Gamma_{jk}(G_i)A^{-1}_{k\mu} = \sum_{j,k}\Gamma_{jk}(G_i)\sum_{\mu} A^{-1}_{k\mu}A_{\mu j} =$$
$$= \sum_{j,k}\Gamma_{jk}(G_i)\delta_{kj} = \sum_{j}\Gamma_{jj}(G_i) = X(G_i),$$

i.e. the character of the element remains the same in all representations. It may easily be demonstrated that the characters of all the elements of one class are the same. Indeed, since the elements of one class are conjugate to one another, for instance,

$$T = VSV^{-1},$$

this means that the representations of the elements are related by means of the similitude transformation

$$\Gamma(T) = \Gamma(VSV^{-1}) = \Gamma(V)\Gamma(S)\Gamma(V^{-1}) = \Gamma(V)\Gamma(S)\Gamma^{-1}(V),$$

whence

$$X(T) = X(S). \tag{A.74}$$

The fundamenlal relation between the elements of irreducible representations enables the relation between the characters of the elements in various irreducible representations to be found as well. Set $\mu = \nu$, $\mu' = \nu'$, and perform summation of the orthogonality relation over μ and μ' to obtain

$$\sum_{\mu,\mu'}\sum_{G_i}\Gamma^{(\alpha)*}_{\mu\mu}(G_i)\Gamma^{(\beta)}_{\mu'\mu'}(G_i) = \sum_{G_i} X^{(\alpha)*}(G_i)X^{(\beta)}(G_i) =$$
$$= \sum_{\mu,\mu'}\frac{g}{f_\alpha}\delta_{\alpha\beta}\delta_{\mu\mu'}\delta_{\mu\mu'} = g\delta_{\alpha\beta}, \tag{A.75}$$

or

$$\sum_{G_i} X^{(\alpha)*}(G_i)X^{(\beta)}(G_i) = g\delta_{\alpha\beta}. \tag{A.76}$$

i.e. the characters of the elements form a system of orthogonal vectors each $\sqrt{g}$ long in the space of the group G.

Instead of the sum over the elements of the group, G_i, one can take the sum over the classes C_i because the characters of the elements of one class are equal.

Denote the number of elements in the class C_i by ρ_i and obtain

$$\sum_{C_i} \sqrt{\frac{\rho_i}{g}} \mathrm{X}^{(\alpha)*}(C_i) \sqrt{\frac{\rho_i}{g}} \mathrm{X}^{(\beta)}(C_i) = \delta_{\alpha\beta}. \qquad (A.77)$$

This equation shows that the characters of irreducible representations form a system of orthogonal vectors in the space of the classes C_i of the group G.

The characters of the element classes in irreducible representations or, as is also said, the characters of irreducible representation, may be found on the basis of some general equations, or particular relations. At present the tables of characters of irreducible representations have been compiled for all the main groups. These tables will be presented below. They are compiled for isomorphic groups shown in the top left corner of the table, usually in the Schoenflies notation. The top line shows the classes and the numbers of elements of each group, and the left column shows the notations of the irreducible representations. The lines of the table indicate the characters of the elements (classes) in specific irreducible representations; the columns of the table indicate the characters of the classes (elements) in different irreducible representations. In addition, the right-hand side of the table usually shows basic functions reflecting symmetry property of functions which are transformed with the aid of irreducible representations shown in the table. This will be the subject of the next section.

The characters of reducible representations may be expressed in terms of the characters of irreducible representations. Suppose the representation is in a reduced form from which it may be seen what irreducible representations $\Gamma^{(\alpha)}(G_i)$ and in what number a_α make up the reducible representation. One may write an obvious expression for the character of the reducible representation $\mathrm{X}(G_i)$:

$$\mathrm{X}(G_i) = \sum_{\alpha=1}^{r} a_\alpha \mathrm{X}^{(\alpha)}(G_i). \qquad (A.78)$$

If the reducible representation is in the non-reduced form, the numbers a_α may not be obtained directly. However, since the character is an invariant of the similitude transformation, the numbers remain the same for all equivalent representations. Symbolically, this is written as follows:

$$\Gamma(G_i) = \sum_{\alpha=1}^{r} a_\alpha \Gamma^{(\alpha)}(G_i). \qquad (A.79)$$

If one finds the numbers a_α this will mean that he has found what irreducible representations make up the specific reducible representation. In this case the reducible representation is said to be expanded in irreducible representations. The numbers a_α may be obtained from the condition of orthogonality of the characters of the irreducible representations.

Multiply the equation

$$X(G_l) = \sum_{\alpha=1}^{r} a_\alpha X^{(\alpha)}(G_l) \tag{A.80}$$

by $X^{(\beta)*}(G_l)$ and perform the summation over the elements of the group to obtain

$$\sum_{G_l} X^{(\beta)*}(G_l)\, X(G_l) = \sum_{\alpha=1}^{r} a_\alpha \sum_{G_l} X^{(\beta)*}(G_l)\, X^{(\alpha)}(G_l) =$$

$$= \sum_{\alpha=1}^{r} a_\alpha g \delta_{\alpha\beta} = g a_{\alpha\beta}, \tag{A.81}$$

i.e.

$$a_\alpha = \frac{1}{g} \sum_{G_l} X^{(\alpha)*}(G_l)\, X(G_l). \tag{A.82}$$

Thus, a_α may be found with the aid of known characters $X^{(\alpha)}(G_l)$ of the irreducible representations and of the character $X(G_l)$ of the given representation.

6. THE BASIS OF A REPRESENTATION

We introduced the representation of a group with the aid of the homomorphism concept. There is, however, another approach to it.

Suppose that f linear-independent orthonormalized functions $\psi_n(\mathbf{r})$ $(n = 1, \ldots, f)$ are specified in a linear space of f-dimensional functions such that

$$\int \psi_n^*(\mathbf{r})\, \psi_m(\mathbf{r})\, d\tau = \delta_{nm}. \tag{A.83}$$

The set of $\psi_n(\mathbf{r})$ may serve as a basis for the expansion of arbitrary functions $\psi(\mathbf{r})$ belonging to the same class.

Recall that the elements of the group G_l are operators $\hat{G}_l$ which are applied to the space and to the functions contained in it:

$$\hat{G}_l \psi_k(\mathbf{r}) = \psi_k(\hat{G}_l \mathbf{r}). \tag{A.84}$$

Expand the functions $\psi_k(\hat{G}_l\mathbf{r})$ in basic functions as follows:

$$\hat{G}_l\psi_k(\mathbf{r})=\psi_k(\hat{G}_l\mathbf{r})=\sum_l \Gamma_{lk}(\hat{G}_l)\,\psi_l(\mathbf{r}). \tag{A.85}$$

As distinct from the usual method of notation, the index of the original function comes second and not first. This is more convenient for future presentation. Premultiplying this equation by $\psi_m^*(\mathbf{r})$ and integrating over the co-ordinates we obtain

$$\int \psi_m^*(\mathbf{r})\,\hat{G}_l\psi_k(\mathbf{r})\,d\tau=(\hat{G}_l)_{mk}=\sum_l \Gamma_{lk}(\hat{G}_l)\,\delta_{lm}=\Gamma_{mk}(\hat{G}_l) \tag{A.86}$$

i.e.

$$\Gamma_{lk}(\hat{G}_l)=\int \psi_l^*(\mathbf{r})\,\hat{G}_l\psi_k(\mathbf{r})\,d\tau. \tag{A.87}$$

It follows that the expansion coefficients of the function $\hat{G}_l\psi_k(\mathbf{r})$ in the basis are simply the matrix elements of the operator computed with the aid of two corresponding basic functions. Form a matrix $\Gamma(\hat{G}_l)$ for the elements $\Gamma_{lk}(\hat{G}_l)$. With its aid one is able to write the expansion of the functions $\hat{G}_l\psi_k^*(\mathbf{r})$ in the basic functions $\{\psi_k(\mathbf{r})\}$. In other words, the matrices $\Gamma(\hat{G}_l)$ show how the basic functions are transformed by each other when acted upon by the group elements:

$$\hat{G}_l\begin{pmatrix}\psi_1(\mathbf{r})\\ \psi_2(\mathbf{r})\\ \vdots\\ \cdot\\ \psi_f(\mathbf{r})\end{pmatrix}=\Gamma(\hat{G}_l)\begin{pmatrix}\psi_1(\mathbf{r})\\ \psi_2(\mathbf{r})\\ \vdots\\ \cdot\\ \psi_f(\mathbf{r})\end{pmatrix}. \tag{A.88}$$

The form of the matrix $\Gamma(\hat{G}_l)$, naturally, depends on the basis selected. Choose another basis $\{\psi_k'(\mathbf{r})\}$ which yields the matrices $\Gamma'(\hat{G}_l)$. Find the relationship between the matrices $\Gamma(\hat{G}_l)$ and $\Gamma'(\hat{G}_l)$. Let the basis $\{\psi_k'(\mathbf{r})\}$ be expressed in the basis $\{\psi_k(\mathbf{r})\}$ with the aid of some matrix $\tilde{A}$:

$$\{\psi_k'(\mathbf{r})\}=\tilde{A}\,\{\psi_k(\mathbf{r})\}, \tag{A.89}$$

or

$$\psi_k'(\mathbf{r})=\sum_j a_{jk}\psi_j(\mathbf{r}). \tag{A.90}$$

It may be shown on the grounds that the "old" and the "new" bases are orthonormalized, that the matrix A is unitary: $A^{-1}=$ $=A^{+}=\tilde{A}^{*}$.

Write

$$\Gamma'_{lm}(\hat{G}_i) = \int \psi_l'^{*}(\mathbf{r})\,\hat{G}_i \psi'_m(\mathbf{r})\,d\tau = \int \sum_n a^{*}_{nl}\psi^{*}_n(\mathbf{r})\,\hat{G}_i \sum_p a_{pm}\psi_p(\mathbf{r})\,d\tau =$$
$$= \sum_{n,\,p} a^{*}_{nl}a_{pm}\Gamma_{np}(\hat{G}_i) = \sum_{n,\,p} \tilde{a}^{*}_{ln}\,\Gamma_{np}(\hat{G}_i)\,a_{pm} = (A^{+}\Gamma(\hat{G}_i)\,A)_{lm}, \quad \text{(A.91)}$$

or

$$\Gamma'(\hat{G}_i) = A^{-1}\Gamma(\hat{G}_i)\,A \quad \text{(A.92)}$$

and

$$\Gamma(\hat{G}_i) = A\Gamma'(\hat{G}_i)\,A^{-1}. \quad \text{(A.93)}$$

Hence, the transition to a new basis results in a similitude transformation of the matrices Γ and Γ'.

Find the relationship between the matrices of the elements $\hat{T}$, $\hat{S}$ and of their product $\hat{T}\hat{S}$:

$$\Gamma_{ij}(\hat{T}\hat{S}) = \int \psi^{*}_i(\mathbf{r})\,\hat{T}\hat{S}\,\psi_j(\mathbf{r})\,d\tau = \int \psi^{*}_i(\mathbf{r})\,\hat{T}\sum_l \Gamma_{lj}(\hat{S})\,\psi_l(\mathbf{r})\,d\tau =$$
$$= \sum_l \Gamma_{lj}(\hat{S})\,\Gamma_{il}(\hat{T}) = \sum_l \Gamma_{il}(\hat{T})\,\Gamma_{lj}(\hat{S}) = (\Gamma(\hat{T})\,\Gamma(\hat{S}))_{ij}, \quad \text{(A.94)}$$

or

$$\Gamma(\hat{T}\hat{S}) = \Gamma(\hat{T})\,\Gamma(\hat{S}).$$

The last relation shows the matrices $\Gamma(\hat{G}_i)$ composed of the matrix elements of the operators $\hat{G}_i$, the place of which is taken by the group elements, to satisfy the conditions imposed on the matrices of the group representation. In other words, the basic functions enable the group representation to be constructed. Note that the choice of the basic functions is, generally, quite arbitrary; at the same time they yield such invariants as the characters of the group elements.

Consider now the problem of the reduction of the representations from the point of view of basic function transformations.

Let $\Gamma^{(\alpha)}(\hat{G}_i)$ be an irreducible group representation induced by a basis $\{\psi_k(\mathbf{r})\}$ of the dimensionality f_α.

The following relation holds:

$$\hat{G}_i \begin{pmatrix} \psi_1(\mathbf{r}) \\ \psi_2(\mathbf{r}) \\ \cdot \\ \cdot \\ \cdot \\ \psi_{f_\alpha}(\mathbf{r}) \end{pmatrix} = \begin{pmatrix} \Gamma_{11}(\hat{G}_i) & \ldots & \Gamma_{f_\alpha 1}(\hat{G}_i) \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \Gamma_{1 f_\alpha}(\hat{G}_i) & \ldots & \Gamma_{f_\alpha f_\alpha}(\hat{G}_i) \end{pmatrix} \begin{pmatrix} \psi_1(\mathbf{r}) \\ \psi_2(\mathbf{r}) \\ \cdot \\ \cdot \\ \cdot \\ \psi_{f_\alpha}(\mathbf{r}) \end{pmatrix}. \quad \text{(A.95)}$$

It shows that the action of the element $\hat{G}_i$ of the group G on the basis $\{\psi_k(\mathbf{r})\}$ results in a mutual transformation of the basic functions so that their *mixing* takes place.

Now let $\Gamma(\hat{G}_i)$ be a reducible representation but of an arbitrary form. Obviously, this case will be undistinguishable from the former since the basic functions will again be subject to mutual transformations. However, everything will look different if the representation is reduced, i.e. it is of a quasidiagonal form:

$$\hat{G}_i \begin{pmatrix} \psi_1(\mathbf{r}) \\ \psi_2(\mathbf{r}) \\ \cdot \\ \cdot \\ \cdot \\ \psi_f(\mathbf{r}) \end{pmatrix} = \begin{pmatrix} \Gamma^{(1)}(\hat{G}_i) & 0 & 0 \\ 0 & \Gamma^{(2)}(\hat{G}_i) & 0 \\ 0 & 0 & \Gamma^{(3)}(\hat{G}_i) \end{pmatrix} \begin{pmatrix} \psi_1(\mathbf{r}) \\ \psi_2(\mathbf{r}) \\ \cdot \\ \cdot \\ \cdot \\ \psi_f(\mathbf{r}) \end{pmatrix}. \quad \text{(A.96)}$$

This expression shows the basis of the linear f-dimensional space to decompose into invariant subspaces corresponding to irreducible representations of the group. When an arbitrary element of the group $\hat{G}_i$ acts on any basic function $\psi_j(\mathbf{r})$, the latter is transformed according to the irreducible representation to which it belongs. The mixing of the basic functions transformed according to different irreducible representations does not take place. This fact is of practical importance since in most cases the solution of quantum-mechanical problems involves the use of equations in matrix form, and they assume the simplest form if calculated with the aid of the basic functions of irreducible representations. The basic functions of the irreducible representations are said to diagonalize the equations in the easiest way.

The basic functions of irreducible representations may be selected arbitrarily, the form of the representation matrices being variable depending on the choice of the functions, but the characters of the elements being the same. Suppose we have to find the characters with the aid of functions satisfying the Schrödinger equation the solutions of which it is difficult or impossible to obtain. In this case we may take for the basis arbitrary orthonormalized functions with the same symmetry properties as the solutions sought and calculate the characters with their aid.

The examples of basic functions are shown in the group character tables.

7. DIRECT PRODUCT OF REPRESENTATIONS

Two kinds of matrix products are considered in the matrix theory. The ordinary matrix product is obtained with the aid of the "line by column" multiplication rule. There is, however, the so-called direct matrix product, as well, which we will consider using two two-row square matrices A and B as an example:

$$A=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad B=\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}. \tag{A.97}$$

A matrix C is termed direct product of the matrix A by the matrix B (denoted $C=A\times B$) if the matrix element of the matrix C is equal to the product of the matrix elements of the multipliers:

$$c_{i,j\,kl}=a_{ik}b_{jl}. \tag{A.98}$$

Two pairs of indices are normally used in conjunction with the matrix C elements composed from the indices of the multipliers. The matrix $A\times B$ may be represented in the form

$$C=A\times B=\begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{pmatrix} \tag{A.99}$$

in full accordance with the definition

$$\{A\times B\}_{ij,\,kl}=a_{ik}b_{jl}. \tag{A.100}$$

Find the trace of the matrix C:

$$\begin{aligned} \operatorname{Sp} C &= \sum_{i,j} c_{ij,\,ij}=a_{11}b_{11}+a_{11}{}'_{22}+a_{22}b_{11}+a_{22}b_{22}= \\ &= a_{11}(b_{11}+b_{22})+a_{22}(b_{11}+b_{22})=(a_{11}+a_{22})(b_{11}+b_{22})= \\ &= \operatorname{Sp} A \operatorname{Sp} B. \end{aligned} \tag{A.101}$$

Hence, the trace of the direct product of matrices is equal to the product of the traces of the multipliers.

Apply this concept to the representation matrices. Let $\Gamma^{(1)}(\hat{G}_l)$ and $\Gamma^{(2)}(\hat{G}_i)$ be two representations of dimensionalities f_1 and f_2, respectively. Form a direct product of the matrices:

$$\Gamma(\hat{G}_l)=\Gamma^{(1)}(\hat{G}_l)\times\Gamma^{(2)}(\hat{G}_l). \tag{A.102}$$

It may be demonstrated that

$$\begin{aligned} \Gamma(\hat{G}_l\hat{G}_k) &= \Gamma^{(1)}(\hat{G}_i\hat{G}_k)\times\Gamma^{(2)}(\hat{G}_l\hat{G}_k)= \\ &= \Gamma^{(1)}(\hat{G}_i)\Gamma^{(1)}(\hat{G}_k)\times\Gamma^{(2)}(\hat{G}_i)\Gamma^{(2)}(\hat{G}_k)= \\ &= (\Gamma^{(1)}(\hat{G}_i)\times\Gamma^{(2)}(\hat{G}_l))(\Gamma^{(1)}(\hat{G}_k)\times\Gamma^{(2)}(\hat{G}_k)), \end{aligned} \tag{A.103}$$

i.e. that the direct product is a group representation too.

If the multipliers $\Gamma^{(1)}(\hat{G}_l)$ and $\Gamma^{(2)}(\hat{G}_l)$ are irreducible representations, their direct product will, generally, be a reducible representation. This may be proved in the following way. Let the greatest dimensionalities of $\Gamma^{(1)}(\hat{G}_l)$ and $\Gamma^{(2)}(\hat{G}_l)$ be $f_1=3$ and $f_2=3$. Then the dimensionality of the direct product will be $f=f_1 \cdot f_2=9$ and, consequently, it should be reducible.

The expansion of a direct product into irreducible representations is accomplished on the basis of the theorem about the orthogonality of the characters of irreducible representations:

$$\Gamma^{(1)}(\hat{G}_l)\times\Gamma^{(2)}(\hat{G}_l)=\sum_{\alpha} a_\alpha \Gamma^{(\alpha)}(\hat{G}_l),$$

$$a_\alpha=\frac{1}{g}\sum_{G_l} X^{(\alpha)*}(\hat{G}_l)\, X(\hat{G}_l), \qquad \text{(A.104)}$$

where $X(\hat{G}_l)$ is obtained from the relation

$$X(\hat{G}_l)=X^{(1)}(\hat{G}_l)X^{(2)}(\hat{G}_l). \qquad \text{(A.105)}$$

The following relation may be written for the irreducible representations:

$$\Gamma^{(\alpha)}(\hat{G}_l)\times\Gamma^{(\beta)}(\hat{G}_l)=\sum a_{\alpha\beta\delta}\Gamma^{(\delta)}(\hat{G}_l),$$

$$a_{\alpha\beta\delta}=\frac{1}{g}\sum_{G_l} X^{(\delta)*}(\hat{G}_l)\, X^{(\alpha)}(\hat{G}_l)\, X^{(\beta)}(\hat{G}_l). \qquad \text{(A.106)}$$

In Sec. 3 we introduced the concept of the direct product of groups: a group G is termed direct product of the groups $G^{(1)}$ and $G^{(2)}$ if the element of the group G is a two-by-two product of the elements of the groups being multiplied, $G_l^{(1)}G_k^{(2)}$. If r_1 and r_2 are numbers of classes in the groups, their product will contain $r=r_1r_2$ classes, and for this reason the number of irreducible representations of the group $G^{(1)}\times G^{(2)}$ will be r_1r_2.

The theory of groups contains proof of the proposition that the direct product of irreducible representations $\Gamma^{(\alpha)}(\hat{G}_l^{(1)})\times\Gamma^{(\beta)}(\hat{G}_k^{(2)})$ is an irreducible representation of the direct product of groups $G^{(1)}\times G^{(2)}$. Omitting the proof of the theorem we shall confine ourselves to showing that all the irreducible representations of the direct group product may be specified in this way. The dimensionality of the direct product is the product of the dimensionalities of the multipliers, i.e. $f_\alpha^{(1)}f_\beta^{(2)}$. Find the sum of the squares of the dimensionalities:

$$\sum_{\alpha,\,\beta} f_\alpha^{(1)^2} f_\beta^{(1)^2}=\Big(\sum_{\alpha} f_\alpha^{(1)^2}\Big)\Big(\sum_{\beta} f_\beta^{(2)^2}\Big)=g_1g_2=g, \qquad \text{(A.107)}$$

i.e., indeed, the sum of the squares of the dimensionalities of the irreducible representation of a group is equal to its order.

The characters of the elements of the groups in the irreducible representations are equal to the product of the characters' of the multipliers.

We introduced the concept of the direct product of representations again resorting to matrix properties. Consider this concept in connection with the properties of basic functions.

Suppose we have two bases, $\{\psi_k(\mathbf{r})\}$ and $\{\varphi_j(\mathbf{r})\}$, which are transformed with the aid of representations $\Gamma^{(1)}(\hat{G}_i)$ and $\Gamma^{(2)}(\hat{G}_i)$, respectively.

Consider the product of two basic functions $\psi_k(\mathbf{r})\varphi_j(\mathbf{r})$. Apply to this function the operator G_i belonging to the group G:

$$\begin{aligned}\hat{G}_i\psi_k(\mathbf{r})\varphi_j(\mathbf{r}) &= \psi_k(\hat{G}_i\mathbf{r})\varphi_j(G_i\mathbf{r}) = \\ &= \left\{\sum_l \Gamma^{(1)}_{lk}(\hat{G}_i)\quad {}_l(\mathbf{r})\right\}\left\{\sum_m \Gamma^{(2)}_{mj}(\hat{G}_i)\varphi_m(\mathbf{r})\right\} = \\ &= \sum_{l,\, m} \Gamma^{(1)}_{lk}(\hat{G}_i)\Gamma^{(2)}_{mj}\psi_l(\mathbf{r})\varphi_m(\mathbf{r}). \qquad (A.108)\end{aligned}$$

This relation may be interpreted as follows. The functions $\psi_k(\mathbf{r})\varphi_j^{(\mathbf{r})}$ form a basis of the dimensionality f_1f_2 which is transformed with the aid of the representation $\Gamma(\hat{G}_i)$ whose matrix element is of the form

$$\Gamma_{lm,\,kj}(\hat{G}_i) = \Gamma^{(1)}_{lk}(\hat{G}_i)\Gamma^{(2)}_{mj}(\hat{G}_i). \qquad (A.109)$$

This agrees with the definition of the matrix element of the direct product of matrices.

In other words, a product of two basic functions, $\psi_k(\mathbf{r})\varphi_j(\mathbf{r})$, is transformed with the aid of the direct product of their representations $\Gamma^{(1)}(\hat{G}_i)\times\Gamma^{(2)}(\hat{G}_i)$. It is quite obvious that should we desire to obtain a representation with the aid of which the product of p basic functions $\psi_j^{(1)}(\mathbf{r})\psi_k^{(2)}(\mathbf{r})\ldots\psi_l^{(p)}(\mathbf{r})$ is transformed, we would obtain it in the form of the direct product of the representations $\Gamma^{(1)}(\hat{G}_i)\times\Gamma^{(2)}(\hat{G}_i)\times\ldots\Gamma^{(p)}(\hat{G}_i)$.

This fact forms the basis of the method of finding "selection rules" by establishing conditions in which a quantity characteristic of some process is either zero or non-zero. If it is zero the corresponding processes are termed forbidden, if it is non-zero the processes are termed allowed. Selection rules for the adsorption and radiation of light may be cited as the best known example.

To illustrate the point let us consider an integral of the form

$$I = \int \varphi^*\hat{L}\psi\,d\tau. \qquad (A.110)$$

If G is a symmetry group of the physical system, the action of the element $\hat{G}_i$ on the system space should result in the quantity

I remaining invariant:

$$\hat{G}_i I = \int \hat{G}_i \varphi^* \hat{L} \psi \, d\tau = \int \varphi^* (\hat{G}_i \mathbf{r}) \, \hat{L} (\hat{G}_i \mathbf{r}) \, \psi (\hat{G}_i \mathbf{r}) \, d\tau'. \qquad \text{(A.111)}$$

Should we regard the functions φ, $\hat{L}$ and ψ as basic functions of some group representations $\Gamma^{(1)}(\hat{G}_i)$, $\Gamma^{(2)}(\hat{G}_i)$ and $\Gamma^{(3)}(\hat{G}_i)$, we would obtain that $\varphi^* \hat{L} \psi$ is transformed with the aid of the representation $\Gamma^{(1)}(\hat{G}_i) \times \Gamma^{(2)}(\hat{G}_i) \times \Gamma^{(3)}(\hat{G}_i)$. Expand it in irreducible representations of the group G. The quantity I will be nonzero only if the direct product contains a unit or trivial representation.

8. POINT GROUPS

We have already discussed some of the simplest point groups. Now we shall consider their representations together with examples of more complex groups.

1. Consider first cyclic groups G formed with the element A:

$$G: A^1, \ A^2, \ \ldots, \ A^{g-1}, \ A^g = E \qquad \text{(A.112)}$$

The order of the element, g, is the same as the order of the group. The number of classes is equal to the number of elements, therefore the number of irreducible representations of the group is equal to the number of elements, g. It follows from the relation

$$\sum_{\alpha=1}^{g} f_\alpha^2 = g \qquad \text{(A.113)}$$

that the irreducible representations must, of necessity, be unidimensional.

Since the representations are unidimensional the table of the characters of irreducible representations contains the irreducible representations themselves. They may be obtained with the aid of the following simple method. Denote the character of the element A in the α-irreducible representation by $X^{(\alpha)}(A)$. The character of the element A^m is $X^{(\alpha)}(A^m)$; however, since in this case the character coincides with the representation, we may write

$$X^{(\alpha)}(A^m) = [X^{(\alpha)}(A)]^m. \qquad \text{(A.114)}$$

Taking into account that $A^g = E$ and that $X^{(\alpha)}(E) = 1$ (a unit element is represented by a unit matrix, and its character is equal to the dimensionality of the representation) we obtain

$$[X^{(\alpha)}(A)]^g = 1. \qquad \text{(A.115)}$$

In order to be able to distinguish between the irreducible representations we write the latter relation in the form

$$[\mathrm{X}^{(\alpha)}(A)]^{g}=1\cdot 1^{(\alpha)};\quad \alpha=1,\,2,\,\ldots,\,g, \tag{A.116}$$

whence

$$\mathrm{X}^{(\alpha)}(A)=\sqrt[g]{1^{\alpha}} \tag{A.117}$$

and

$$\mathrm{X}^{(\alpha)}(A^{m})=\sqrt[g]{1^{m\alpha}}. \tag{A.118}$$

Since

$$\sqrt[g]{1^{\alpha}}=e^{i\frac{2\pi\alpha}{g}}, \tag{A.119}$$

it follows that

$$\mathrm{X}^{(\alpha)}(A)=e^{i\frac{2\pi\alpha}{g}} \tag{A.120}$$

and

$$\mathrm{X}^{(\alpha)}(A^{m})=e^{i\frac{2\pi\alpha m}{g}}. \tag{A.121}$$

Putting $\alpha=1;\ 2,\ \ldots,\ g$ we obtain non-equivalent irreducible representations of a cyclic group. This may be easily checked with the aid of orthogonality theorems. For instance, the sum over the elements of the group is

$$\sum_{m=1}^{2} e^{-i\frac{2\pi\alpha_1 m}{g}}\, e^{i\frac{2\pi\alpha_2 m}{g}}=\sum_{m=1}^{g} e^{i\frac{2\pi(\alpha_2-\alpha_1)}{g}m}=\frac{1-\left(e^{i\frac{2\pi(\alpha_2-\alpha_1)}{g}}\right)^{g}}{1-e^{i\frac{2\pi(\alpha_2-\alpha_1)}{g}}} \tag{A.122}$$

for α_1 and α_2.

For $\alpha_2\neq\alpha_1$ the sum is zero, for $\alpha_2=\alpha_1$ it is equal to g in full agreement with the orthogonality theorems.

2. There are two representations of the isomorphic second-order groups C_2, C_s and C_i which we may find putting $\alpha=1,2$:

C_2	E	C_2
C_s	E	σ
C_i	E	i
Γ_1	1	1
Γ_2	1	-1

(A.123)

It may be seen from the table of characters that there are two types of basic functions. The function ψ_g which is transformed with the aid of the irreducible representation Γ_1 $(\alpha=2)$ remains unchanged; it is termed *symmetrical*, or *even*. The function ψ_u which is transformed with the aid of the irreducible representation $\Gamma_2 (\alpha = 1)$ changes sign when acted upon by a non-identical group element; it is termed *antisymmetrical*, or *odd*. It must, however, be kept in mind that the nature of non-parity of the basic functions is different in different groups. The "parity" in co-ordinates which follows from the relations $\hat{C}_2\psi_u(\mathbf{r}) = (-\psi_u(\mathbf{r}))$; $\hat{\sigma}\psi_u(\mathbf{r})$ and $\hat{i}\psi_u(\mathbf{r})$ is different, for instance,

$$\hat{i}\psi_u(\mathbf{r}) = \psi_u(\hat{i}\mathbf{r}) = \psi_u(-x, -y, -z) = -\psi_u(x, y, z) = -\psi_u(\mathbf{r});$$

$$\hat{C}_2^{(z)}\psi_u(\mathbf{r}) = \psi_u(\hat{C}_2^{(z)}\mathbf{r}) = \psi_u(-x, -y, z) = -\psi_u(x, y, z) = -\psi_u(\mathbf{r}); \tag{A.124}$$

$$\hat{\sigma}_z\psi_u(\mathbf{r}) = \psi_u(\hat{\sigma}_z\mathbf{r}) = \psi_u(x, y, -z) = -\psi_u(x, y, z) = -\psi_u(\mathbf{r}).$$

3. Groups C_6 and C_{6v}. Consider now a more complex group. The group C_6 is a group of rotations of a hexagon or a hexagonal prism. It is a cyclic group of the order $g = 6$. Its six irreducible unidimensional representations are

$$\Gamma^{(\alpha)}(C_6^l) = e^{i\frac{2\pi\alpha}{6}l}; \quad \alpha = 1, 2, \ldots, 6, \quad l = 1, 2, \ldots, 6. \tag{A.125}$$

The group C_{6v} may be obtained from the group C_6 by the addition of one vertical plane σ_v. It follows automatically from the existence of a C_6-axis that the system must possess five additional vertical planes. Thus, we have twelve elements: C_6, $C_6^2 = C_3$; $C_6^3 = C_2$; $C_6^4 = C_3^2 = C_3^{-1}$; $C_6^5 = C_6^{-1}$ and $C_6^6 = E$, and six planes σ_v. Consider the products of the elements to discover possible new elements. The product of two planes σ_v' and σ_v'' is the rotation through the angle equal to twice the angle between the planes. If we take into account that the minimum angle between two neighbouring planes is $\frac{2\pi}{12}$ and the double angle is exactly $\frac{2\pi}{6}$, it will be clear that the reflection from two arbitrary planes will result in the rotation through an angle multiple of $\frac{2\pi}{6}$, i.e. will yield an element of the type C_6^l. Now consider the product of the elements of the form $C_6^l\sigma_v$. It may easily be seen that this yields either the operation σ_v' or $C_6^{l'}$, i.e. again does not constitute a new element. Hence, the group C_{6v} order is $g = 12$.

Find the number of classes. Each element C_6^l in the group C_6 forms its own class. Evidently, the elements C_3 and C_2 of different orders cannot belong to the same class. We ought to check the possibility of the elements C_3 and C_3^{-1}, C_6 and C_6^{-1} being conjugate, because the group C_{6v} contains elements σ_v that do not commute with the axes. Consider the element $\sigma_v C_6^l \sigma_v^{-1} = \sigma_v C_6^l \sigma_v$. It may be shown analytically or geometrically that $\sigma_v C_6^l \sigma_v^{-1} = C_6^{-l}$. For this reason the elements C_6^l and C_6^{-l} belong to the same class and, by force of this, the rotations C_6^l constitute the classes E; C_6; C_6^{-1}; C_3; C_3^{-1}; C_2. It follows from $\sigma_v C_6^l \sigma_v^{-1} = C_6^{-l}$ that $C_6^l \sigma_v = \sigma_v C_6^{-l}$, therefore $C_6^l \sigma_v C_6^{-l} = C_6^{2l}$. Now consider an element of the form $\sigma_v' \sigma_v \sigma_v'$. It may be written in the following form: $\sigma_v' \sigma_v \sigma_v \sigma_v \sigma_v' = C_{12}^m \sigma_v C_{12}^{-m} = C_6^m$. This result means that a triple reflection in two planes, σ_v', σ_v and σ_v', rotates the plane σ_v through an angle multiple of $\frac{2\pi}{6}$. For this reason it is impossible to make all the planes coincide with each other, this being possible only for every second plane. In other words, there are two classes of conjugate planes, and by force of this the group C_{6v} contains six classes: E; $2C_6$; $2C_3$; C_2; $3\sigma_v'$; $3\sigma_v''$. The equation

$$\sum_{\alpha=1}^{6} f_\alpha^2 = 12 \tag{A.126}$$

is satisfied only for $f_\alpha = 1, 1, 1, 1, 2, 2$, i.e. the group C_{6v} contains four one-dimensional and two two-dimensional irreducible representations.

Table of irreducible representations of the group C_{6v}

C_{6v}	E	$C_2^{(z)}$	$2C_3^{(z)}$	$2C_6^{(z)}$	$3\sigma_x$	$3\sigma_y$
Γ_1	1	1	1	1	1	1
Γ_2	1	1	1	1	−1	−1
Γ_3	1	−1	1	−1	1	−1
Γ_4	1	−1	1	−1	−1	1
Γ_5	2	−2	−1	1	0	0
Γ_6	2	2	−1	−1	0	0

4. Group T. The group of tetrahedron axes is denoted by T, or 28. It contains 12 elements: E, $3C_2$, $4C_3$ and $4C_3^2$ written in the order of classes. The existence of four classes is obvious because the elements C_3 and $C_3^{-1} = C_3^2$ are not conjugate. The dimensionalities of the irreducible representations may be found from the equation

$$\sum_{\alpha=1}^{4} f_\alpha^2 = 12,$$

whose solution is $1^2 + 1^2 + 1^2 + 3^2 = 12$. The table of characters is of the following form:

Table of characters of the group T

	T	E	$3C_2$	$4C_3$	$4C_3^2$
Γ_1	A	1	1	1	1
Γ_2	E	1	1	ω	ω^2
Γ_3		1	1	ω^2	ω
Γ_4	F	3	-1	0	0

$$\omega = e^{i\frac{2\pi}{3}}$$

5. The groups T_d and O. The group O, or 432, is a cubic-symmetry group. It contains three four-fold, four three-fold, and six two-fold axes. Altogether there are 24 elements in the group distributed among five classes: $3C_4$, $3C_4^2$, $3C_4^3$, $4C_3$, $6C_2^3$ and E. The table of characters is of the following form:

Table of characters of the group O and T_d

O	T_d	E E	$8C_3$ $8C_3$	$3C_2$ $3C_2$	$6C_2$ $6\sigma_d$	$6C_4$ $6S_4$
Γ_1	A_1	1	1	1	1	1
Γ_2	A_2	1	1	1	-1	-1
Γ_3	E	2	-1	2	0	0
Γ_4	F_1	3	0	-1	-1	1
Γ_5	F_2	3	0	-1	1	-1

The full summetry group of the tetrahedron T_d, or $43m_d$, is isomorphous with the group O. The group T_d contains diagonal planes σ_d, three rotary-reflection axes S_4 lying in these planes, and four three-fold axes. The elements of the group are: E, $3S_4$, $3S_4^2$, $3S_4^2 = 3C_2$, $3S_4^3$, $4C_3$, $4C_3^2$ and $6\sigma_d$, i.e. 24 elements in all distributed among 5 classes.

6. The groups T_h and O_h.

Consider two groups T_h and O_h as an example of a direct product. The group O_h is defined as a direct product of the groups O and C_i.

$$O_h = O \times C_i \tag{A.127}$$

Since group O contains 24 elements distributed among five classes, and group C_i has two elements E and i, the group O_h contains 48 elements distributed among 10 classes. Five classes of the group O_h coincide with the five classes of the group O because they correspond to the product $O \times E$, the five other classes correspond to the product $O \times i$, pure inversion $E_i = i$ and different reflection planes being among them. The group O_h is a full cubic-symmetry group. The table of characters of irreducible representations of the group O_h may be compiled on the basis of the tables of characters of irreducible representation of groups O and C_i.

Various authors, as a rule, use different notations for irreducible representations. Particularly- for the group O_h is this diversity especially pronounced. The table of characters of the group O_h contains most common designations of the irreducible representations of the group O_h. We would like to remind that unidimensional representations are generally designated by the letter A or B, two-dimensional by the letter E and three-dimensional by the letter T or F. The representations even with respect to the inversion have the index "g", odd representations—the index "u".

Classes					Basic fuctions
$6iC_4$	$6iC_2$	i	$3iC_4^2$	$8iC_3$	
1	1	1	1	1	1
1	1	−1	−1	−1	xyz
−1	−1	1	1	1	$x^4(y^2-z^2)+y^4(z^2-x^2)+z^4(x^2-y^2)$
−1	−1	−1	−1	−1	$xyz\,[x^4(y^2-z^2)+y^4(z^2-x^2)+z^4(x^2-y^2)]$
0	0	2	2	−1	$3x^2-r^2;\ \sqrt{3}(y^2-z^2)$
0	0	−2	−2	+1	$xyz(3x^2-r^2);\ \sqrt{3}\,xyz(y^2-z^2)$
1	1	3	−1	0	$yz;\ zx;\ xy$
1	1	−3	1	0	$x;\ y;\ z$
−1	−1	3	−1	0	$yz(y^2-z^2);\ zx(z^2-x^2);\ xy(x^2-y^2)$
−1	−1	−3	1	0	$x(y^2-x^2);\ y(z^2-x^2);\ z(x^2-y^2)$

Such notations are used, for instance, in the books "*Quantum Mechanics*" by Landau and Lifshitz, and "*Quantum Chemistry*" by Eyring, Walter and Kimball. Howars and Jones use the symbols of atomic hydrogen-like function with identical (or almost identical) symmetry as indices. For example, they denote the representation of a state with the wave function which is transformed with the aid of the representation A_{1g} by Γ_s, because the s-state in the hydrogen atom is spherically symmetrical and, by force of this, invariant in relation to all rotations. The properties of the basic function of the Γ_s representation, for which an arbitrary constant, for instance 1, may be employed, are analogous.

This idea stands out most clearly in the notation of the three-dimensional representation Γ_p which employs for its basic functions the components of the radius-vector **r**: x, y, z, the wave functions of the p-state being of the form $xf(\mathbf{r})$, $yf(\mathbf{r})$, $zf(\mathbf{r})$ where $f(\mathbf{r})$ is a spherically symmetrical function. For this reason the three-dimensional representation Γ_p in the notation H, J or Γ_{12} (reads: gamma one-two) in the notation B, S, W may be regarded as a representation of the p-analogue state. Twin indices in the B, S, W notation of two-dimensional (15) and three-dimensional (25) representations are related to conditions of compatibility.

The group T_h is a direct product of the groups T and C_i:

$$T_h = T \times C_i \tag{A.128}$$

Its properties may be derived from the properties of the groups T and C_i.

And a concluding remark: the sum of characters of the elements of the group of all representations, except the unit representation, in compliance with the condition of orthogonality should be zero because this sum may be represented as a scalar product of the characters of a specific representation and of the unit representation. The same sum for the unit representation is equal to the vector length and, hence, to the group order.

9. TRANSLATIONAL GROUPS. BRILLOUIN ZONES

In this paragraph we shall introduce the already familiar concepts of the Brillouin zones, of Bloch functions, etc., but from a different stand-point.

Denote three basic vectors which determine for integers n_1, n_2, n_3 the translation vector

$$\mathbf{n} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3 \tag{A.129}$$

by $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$.

The translation vector determines the translation operator

$$\hat{\mathrm{T}}(\mathbf{n}) = e^{i(\mathbf{n}\nabla)} \tag{A.130}$$

which constitutes the translational group $\hat{\mathrm{T}}(\mathbf{n})$, or $(E/\mathbf{n})$ in Seitz's notation.

It may easily be demonstrated that the translational group $\hat{\mathrm{T}}(\mathbf{n})$ is a direct product of three cyclic groups of translations along the three basic vectors $\mathbf{a}_j$. Indeed, it follows from the obvious condition

$$\hat{\mathrm{T}}^2(\mathbf{n}) = \hat{\mathrm{T}}(\mathbf{n})\,\hat{\mathrm{T}}(\mathbf{n}) = \hat{\mathrm{T}}(2\mathbf{n}) \tag{A.131}$$

that

$$\hat{\mathrm{T}}^l(\mathbf{n}) = \hat{\mathrm{T}}(l\mathbf{n}). \tag{A.132}$$

Now put $n_2 = n_3 = 0$ and obtain

$$\mathbf{n} = n_1\mathbf{a}_1 \tag{A.133}$$

and

$$\hat{\mathrm{T}}(n_1\mathbf{a}_1) = \hat{\mathrm{T}}^{n_1}(\mathbf{a}_1). \tag{A.134}$$

In the same way we may write for $n_2 \neq 0$, $n_3 \neq 0$:

$$\hat{\mathrm{T}}(n_2\mathbf{a}_2) = \hat{\mathrm{T}}^{n_2}(\mathbf{a}_2);$$
$$\hat{\mathrm{T}}(n_3\mathbf{a}_3) = \hat{\mathrm{T}}^{n_3}(\mathbf{a}_3), \tag{A.135}$$

i.e. we may consider three groups of translations along the basic vectors. Since the translations along different basic vectors are independent and since the translation by the vector $\mathbf{a}_1 + \mathbf{a}_2$ may be represented as a consecutive translation by $\mathbf{a}_1$ followed by translation by $\mathbf{a}_2$ (or vice versa, first by $\mathbf{a}_2$, then by $\mathbf{a}_1$), we may write from the point of view of group notations

$$\hat{\mathrm{T}}(\mathbf{a}_1 + \mathbf{a}_2) = \hat{\mathrm{T}}(\mathbf{a}_1) \times \hat{\mathrm{T}}(\mathbf{a}_2) \tag{A.136}$$

with the immediate result

$$\hat{\mathrm{T}}(\mathbf{n}) = \hat{\mathrm{T}}^{n_1}(\mathbf{a}_1) \times \hat{\mathrm{T}}^{n_2}(\mathbf{a}_2) \times \hat{\mathrm{T}}^{n_3}(\mathbf{a}_3). \tag{A.137}$$

In this case a cyclic group of infinite order is transformed into a group of finite order $g = N_1N_2N_3 = g_1g_2g_3$ by the introduction of the Born-Carman cyclic boundary conditions:

$$\hat{\mathrm{T}}^{N_j}(\mathbf{a}_j) = E. \tag{A.138}$$

The irreducible representations of cyclic groups are unidimensional and are specified in the form

$$\Gamma^{(\alpha)}(\hat{\mathrm{T}}(\mathbf{a}_j)) = e^{i\frac{2\pi}{N_j}\alpha_j}, \quad \alpha_j = 0, 1, \ldots, N_j - 1, \tag{A.139}$$

where α_j denotes the "number" of the irreducible representation which forms the element $\hat{\mathrm{T}}(\mathbf{a}_j)$. For an arbitrary element $\hat{T}(n_j\mathbf{a}_j)$ we obtain

$$\Gamma^{(\alpha)}(\hat{\mathrm{T}}(n_j\mathbf{a}_j)) = [\Gamma^{(\alpha)}(\hat{\mathrm{T}}(\mathbf{a}_j))]^{n_j} = e^{i\frac{2\pi\alpha_j n_j}{N_j}}. \tag{A.140}$$

Taking into account that irreducible representations of the direct product of groups is a direct product of the irreducible representations of the multipliers, we may write

$$\Gamma^{(\alpha)}(\hat{\mathrm{T}}(\mathbf{n})) = \Gamma^{(\alpha_1)}(\hat{\mathrm{T}}(n_1\mathbf{a}_1)) \times \Gamma^{(\alpha_2)}(\hat{\mathrm{T}}(n_2\mathbf{a}_2)) \times \Gamma^{(\alpha_3)}(\hat{\mathrm{T}}(n_3\mathbf{a}_3)) =$$
$$= e^{i\left(\frac{2\pi}{N_1}\alpha_1 n_1 + \frac{2\pi}{N_2}\alpha_2 n_2 + \frac{2\pi}{N_3}\alpha_3 n_3\right)}. \tag{A.141}$$

For the sake of convenience the quantity $\mathbf{n}$ itself may be introduced into the notation of irreducible translations by the vector $\mathbf{n}$. This is conveniently done with the aid of the vector $\mathbf{k}$ built on a basis $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ in the form

$$\mathbf{k} = 2\pi\left(\frac{\alpha_1}{N_1}\mathbf{b}_1 + \frac{\alpha_2}{N_2}\mathbf{b}_2 + \frac{\alpha_3}{N_3}\mathbf{b}_3\right) = (k_1, k_2, k_3). \qquad (A.142)$$

Suppose that the bases $\{\mathbf{a}_j\}$ and $\{\mathbf{b}_l\}$ are mutually orthogonal:

$$(\mathbf{a}_j\mathbf{b}_l) = \delta_{jl}. \qquad (A.143)$$

In th's case the scalar product $(\mathbf{kn})$ assumes the form

$$(\mathbf{kn}) = \left(\sum_{j=1}^{3}\frac{2\pi\alpha_j}{N_j}\mathbf{b}_j, \sum_{l=1}^{3} n_l\mathbf{a}_l\right) = \sum_{j,l}\frac{2\pi\alpha_j n_l}{N_j}(\mathbf{b}_j\mathbf{a}_l) =$$
$$= 2\pi\sum_{j=1}^{3}\frac{\alpha_j}{N_j}n_j = 2\pi\left(\frac{\alpha_1 n_1}{N_1} + \frac{\alpha_2 n_2}{N_2} + \frac{\alpha_3 n_3}{N_3}\right), \qquad (A.144)$$

which enables the irreducible representation of the translational group to be written in a concise form:

$$\Gamma^{(\alpha)}(\hat{T}(\mathbf{n})) = \Gamma^{(\mathbf{k})}(\mathbf{n}) = e^{i(\mathbf{kn})}. \qquad (A.145)$$

For this reason the basic functions, when acted upon by the translational operator, should be transformed with the aid of the irreducible representations $e^{i(\mathbf{kn})}$

$$\hat{T}(\mathbf{n})\psi(\mathbf{r}) = \psi(\mathbf{r}+\mathbf{n}) = e^{i(\mathbf{kn})}\psi(\mathbf{r}). \qquad (A.146)$$

The brief notation for the basic function of the irreducible representation $e^{i(\mathbf{kn})}$ is $\psi_{\mathbf{k}}(\mathbf{r})$.

The number $N_1N_2N_3$ may be regarded as the number of basic cells built on the basis $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$. The number of non-equivalent irreducible representations is exactly $N_1N_2N_3$. They are commonly regarded as points in the $\mathbf{k}$-space, or in the reciprocal space; $\mathbf{k}$-space, being the image of the crystal space in the reciprocal space, is fully determined by the reciprocal lattice and the elementary cell of the reciprocal lattice.

The basis of the reciprocal lattice $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ may be found from the condition $(\mathbf{a}_j\mathbf{b}_l) = \delta_{lj}$ in the form

$$\mathbf{b}_1 = \frac{[\mathbf{a}_2\mathbf{a}_3]}{(\mathbf{a}_1[\mathbf{a}_2\mathbf{a}_3])}; \quad \mathbf{b}_2 = \frac{[\mathbf{a}_1\mathbf{a}_3]}{(\mathbf{a}_1[\mathbf{a}_2\mathbf{a}_3])}; \quad \mathbf{b}_3 = \frac{[\mathbf{a}_1\mathbf{a}_2]}{(\mathbf{a}_1[\mathbf{a}_2\mathbf{a}_3])}. \qquad (A.147)$$

The modulus of the base vector of the reciprocal lattice is

$$|\mathbf{b}_j| = \frac{1}{|\mathbf{a}_j|}, \qquad (A.148)$$

the relationship between the volume of the parallelepiped V_b built on the basis $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ and the volume of the parallelepiped $V_a = (\mathbf{a}_1 [\mathbf{a}_2 \mathbf{a}_3])$ being connected by the relation

$$V_b = (\mathbf{b}_1 [\mathbf{b}_2 \mathbf{b}_3]); \quad V_b = \frac{1}{V_a}. \qquad \text{(A.149)}$$

The projections of the vector **k** on the basis may be expressed in terms of $|\mathbf{a}_j|$:

$$k_j = 2\pi \frac{\alpha_j}{N_j} \frac{1}{a_j} = \frac{2\pi}{L_j} \alpha_j \qquad \text{(A.150)}$$

where $L_j = N_j a_j$ is the length of the crystal edge, and α_j are integers. For a "large" crystal k_j assume closely spaced values, therefore the vector **k** may be regarded as a quasicontinuous quantity. The dimensionality of **k** being $[L^{-1}]$, it became known as the wave vector.

The basic function $\psi_\mathbf{k}$ may be taken in the form $e^{i(\mathbf{kr})}$. Indeed, the function $e^{i(\mathbf{kr})}$ induces the representation $e^{i(\mathbf{kn})}$:

$$\hat{\mathrm{T}}(\mathbf{n}) e^{i(\mathbf{kr})} = e^{i(\mathbf{k}, \mathbf{r}+\mathbf{n})} = e^{i(\mathbf{kn})} e^{i(\mathbf{kr})}. \qquad \text{(A.151)}$$

Recalling that the basic functions $\psi_\mathbf{k}(\mathbf{r})$ may, generally, be chosen at will we may multiply $e^{i(\mathbf{kr})}$ by an arbitrary periodic function $\varphi_\mathbf{k}(\mathbf{r}+\mathbf{n}) = \varphi_\mathbf{k}(\mathbf{r})$, putting

$$\psi_\mathbf{k}(\mathbf{r}) = e^{i(\mathbf{kr})} \varphi_\mathbf{k}(\mathbf{r}). \qquad \text{(A.152)}$$

The function $e^{i(\mathbf{kr})} \varphi_\mathbf{k}(\mathbf{r})$ results in the same irreducible representations. The function $\psi_\mathbf{k}(\mathbf{r}) = e^{i(\mathbf{kr})} \varphi_\mathbf{k}(\mathbf{r})$ is termed *Bloch function*, or *Bloch wave*; it plays a major part in the solid-state theory.

The area of the reciprocal space containing all the non-equivalent irreducible representations of the translational group $\hat{\mathrm{T}}(\mathbf{n})$ of the crystal is termed *Brillouin zone*. Another name for it is *elementary cell of the reciprocal lattice*. It should, however, be kept in mind that its analogue is not the elementary cell built on the basis $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, but the Wigner-Seitz cell.

By definition, the Brillouin zone contains all the non-equivalent representation of the translational group. Equivalent irreducible representations may be located only on the zone boundaries. This enables the following method of constructing the zone to be used. Take a site of the reciprocal lattice as the origin and draw segments connecting it with several neighbouring sites. Set up orthogonal planes intersecting the segments in the middle. The Brillouin zone will be represented by the minimum-volume polyhedron containing the origin of co-ordinates. Indeed, for any crystallographic direction of the reciprocal lattice the points of the zone boundary are separated by the distance not greater than the minimum reciprocal lattice vector for the corresponding direction. The number of faces is determined by the number of sites inside one or two co-ordination spheres. The analytical expression for the boundary planes is

$$(\mathbf{b}, \mathbf{k}+\pi\mathbf{b})=0, \qquad (A.153)$$

where the wave vector determines the state on the boundary plane, and b is the reciprocal lattice vector.

It follows from the definition of the reciprocal lattice basis that the reciprocal lattice for cubic and hexagonal crystals, which are of primary interest to us, is also cubic and hexagonal, respectively. To check this one should introduce a rectangular co-ordinate system with the unit vectors **i**, **j**, **k**, express in it the base vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ and find $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$. Leaving out the computations we present the result: the reciprocal lattice corresponding to the face-centered cubic lattice (fcc) is volume-centered cubic (vcc), and to the volume-centered cubic lattice—face-centered cubic. The reciprocal lattice corresponding to a hexagonal lattice with the parameter ratio $c/a > 1$ is also hexagonal, but with the ratio of these parameters less than unity.

The point symmetry groups of direct and reciprocal lattices are the same. For the hexagonal lattice this is the C_{6v} group, for the cubic lattice—one of the cubic groups T, T_d, T_h, O or O_h.

The Brillouin zone for a hexagonal crystal is a part of a hexagonal prism. The Brillouin zone for vcc lattice is a rhombododecahedron, for the fcc lattice—a polyhedron of 14 faces which may be obtained out of a cube with truncated apexes. The intersecting planes are perpendicular to the space diagonals of the cube and intersect them at 1/4 of their length. Accordingly, they intersect the edges of the cube at distances of 3/4 periods from the cube apexes, the resulting faces having the shape of hexagons (eight) and squares (six).

The diamond lattice may be regarded as consisting of two face-centered cubic lattices one displaced relative to the other along the space diagonal to a distance equal to one quarter of its length. The elementary cell of the diamond cubic lattice has no inversion centre, therefore its symmetry may be described by the group O. The space group of the diamond lattice contains, however, besides simple translations by the lattice period also a glide plane. This transformation may be represented as follows. Place the co-ordinate origin in an apex of the elementary cell. Set up a plane orthogonal to the x-axis and intersecting it at the point $(a/8, 0, 0)$. Translate the lattice by $(\mathbf{a}_2+\mathbf{a}_3)/4$, and next reflect the space in the plane $\sigma_x\,(a/8, 0, 0)$. Find the relationship between the co-ordinates of the original and the transformed space. The point M with the co-ordinates (x, y, z) is initially transformed into the point $(x, y+a/4, z)$, then into the point $(x, y+a/4, z+a/4)$. Should the reflection be from the plane $\sigma_x(0, 0, 0)$, the co-ordinates of the point would change to $(-x, y+a/4, z+a/4)$.

Figure 144 shows the reflection in the plane $\sigma_x(x_0, 0, 0)$ containing the point $(x_0, 0, 0)$. It may be seen from the figure that reflection in the plane $\sigma_x(x_0, 0, 0)$ transforms the point with the co-ordinates (x, y, z) into the point with the co-ordinates $(-x+2x_0, 0, 0)$, this being determined by the condition of reflection $x-x_0=$ $=x_0-x'$, whence $x'=2x_0-x$. On this account the point $(x, y+a/4, z+a/4)$ is transformed in the course of reflection into the point $(-x+a/4, y+a/4, z+a/4)$. But this transformation may be regarded as a reflection in the co-ordinate plane σ_x followed by translation along the space diagonal to a distance equal to a quarter of its length. As was mentioned above, the translation along the space diagonal transforms one face-centered cubic sublattice into the other.

By analogy, one may write the transformation of the other two co-ordinates in the form

$$\left(x+\frac{a}{4},\ -y+\frac{a}{4},\ z+\frac{a}{4}\right),\quad \left(x+\frac{a}{4},\ y+\frac{a}{4},\ -z+\frac{a}{4}\right). \qquad (A.154)$$

Hence, we may say that the combination of translations by the basic vectors with subsequent reflection in the planes $\sigma_x\left(\frac{a}{8}, 0, 0\right)$,

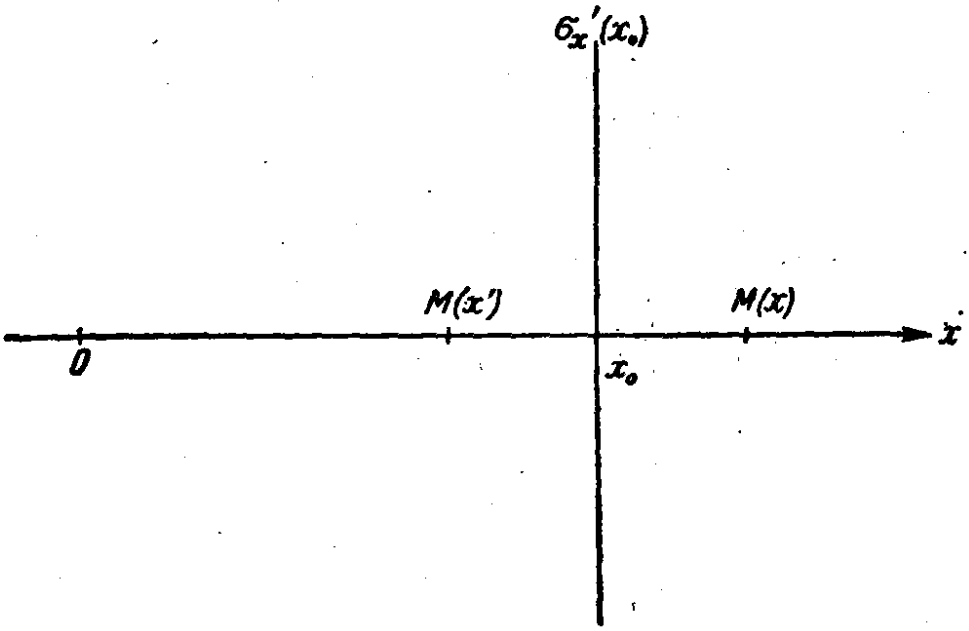


Fig. 144. Reflection in a plane passing through a given point

$\sigma_y\left(0, \frac{a}{8}, 0\right)$ and $\sigma_z\left(0, 0, \frac{a}{8}\right)$ is equivalent to reflection in the co-ordinate planes $\sigma_x, \sigma_y, \sigma_z$ followed by displacement by a quarter of the space diagonal $\hat{T}\left(\frac{d}{4}\right)$.

Consider the product of the operators of the form

$$\hat{\mathrm{T}}\left(\frac{d}{4}\right)\sigma_z\hat{\mathrm{T}}\left(\frac{d}{4}\right)\sigma_y\hat{\mathrm{T}}\left(\frac{d}{4}\right)\sigma_x \tag{A.155}$$

and apply it to the point $M(x, y, z)$. After completing the specified operations we will find that the point $M(\mathbf{r})$ will be transformed into $M(\mathbf{r}')$ where

$$\mathbf{r}' = \left(-x+\frac{3a}{4},\ -y+\frac{a}{4},\ -z-\frac{a}{4}\right). \tag{A.156}$$

But this transformation may be obtained with the aid of the operators $\hat{\mathrm{T}}(a/2,\ 0,\ -a/2)\,\hat{\mathrm{T}}(d/4)\,\hat{\mathrm{i}}$, as well.

10. THE WAVE VECTOR GROUP

The symmetry of the reciprocal lattice is closely related to the symmetry of the direct lattice. The transformations defined by the group G in the space of the crystal may be applied to the reciprocal lattice as well. If $\mathbf{k}$ is a radius vector in the reciprocal lattice space, and $\hat{\mathrm{G}}_l$, an element of the group of point transformations, $\mathbf{k}' = \hat{\mathrm{G}}_l\mathbf{k}$ will yield a group of point transformations in the Brillouin zone.

The centre of the Brillouin zone $\mathbf{k} = (0,0,0)$ is usually chosen as the point relative to which the point transformations are carried out. It is designated by the Greek capital letter Γ. If the space group of the crystal contains rotation, inversion, and reflection planes the translational group and the point transformation group will be subgroups of the space group. In this case the irreducible representations of the point group are termed irreducible representations of the point Γ.

If the space group contains screw axes or reflection planes the method of constructing the point group of the Brillouin zone will be somewhat different. The point group of the Brillouin zone is a group isomorphous with the factor-group of the space group with respect to the pure translations subgroup.

This is the reason why the point group of the Brillouin zone of the diamond lattice is not the cubic axes symmetry group O, but the full cubic-symmetry group $O_h = O \times C_i$ although neither the point group nor the space group of the diamond lattice contains inversion.

The symbol of the diamond lattice space group is $Fd3m$, or O_h^7. It was demonstrated above that not inversion but inversion followed by translation by a quarter of the space diagonal is its symmetry element. The inversion relative to one of the sites results in invariance of only one of the two sublattices. The second sublattice is not invariant with respect to inversion since the inversion results in vacant octants coinciding with the occupied.

In the sphalerite lattice the sublattices are occupied by different atoms; for this reason its space group does not contain a glide plane, and the point symmetry group of the Brillouin zone coincides with the point group of the direct lattice. Recall that the notation of the sphalerite lattice space group is $F43m$, or T_d^1. The point group of the Brillouin zone is the full symmetry group of the tetrahedron T_d.

The designation of the wurtzite lattice space group is $P6mC$, where P denotes a simple (primitive) lattice, and C is the glide plane.

The point **k** is termed *common point of the Brillouin zone* if when acted upon by all the elements of the point group G it yields g points of the Brillouin zone. Since in this case the transformations are performed about the point Γ it is said that the symmetry group G of the Brillouin zone is the symmetry of the point Γ.

The point **k** is termed *symmetry point* as distinct from the common point if it lies on some symmetry element. It is fairly obvious that in all transformations $\hat{G}_i\mathbf{k}$ the number of different points will be below the group order. Their points of symmetry, moreover, will again be transformed into symmetry points.

Introduce the concept of the **wave vector group** $G_\mathbf{k}$: *the totality of transformations* $\hat{G}_\mathbf{k}$ ***which leaves the vector* k *invariant is termed wave vector group.*** Since the transformations in question are part of the point group G, the wave vector group $\hat{G}_\mathbf{k}$ is a subgroup of the point group.

The irreducible representations of the subgroup differ from those of the point group. This point may be illustrated as follows. Suppose the group G has an irreducible representation of maximum dimensionality $f_\alpha = 3$. The order of the group $g > 10$. Let the order of the subgroup be $h < 9$. The irreducible representations of the subgroup cannot have dimensionality above 2; therefore, one irreducible representation of the group should be reducible. A reduction in the subgroup is said to take place. Irreducible representations of the wave vector subgroup are termed *minor representations*. Their form depends on the symmetry points.

The simplest case of a wave vector group is the group of the vector $\mathbf{k} = (0, 0, 0)$, i.e. of the point Γ: all transformations $\hat{G}_i$ belonging to the point group G leave the point Γ invariant. This means that the point group is itself a group of the point Γ: $G_\mathbf{k} = G$. The case when **k** is changed by a vector of the reciprocal lattice $2\pi\mathbf{b}$ should, too, be regarded as one when it remains invariant. Hence, the wave vector group is determined by the equation

$$\hat{G}_k \mathbf{k} = \mathbf{k} + 2\pi \mathbf{b}. \qquad (A.157)$$

Taking the vector **k** at the common point of the Brillouin zone we may obtain from it g vectors of the type $\mathbf{k}' = \hat{G}_i \mathbf{k}$. The set of various vectors $\mathbf{k}' = \hat{G}_i \mathbf{k}$ is termed *star of the vector* **k**. The wave vector group at the common point consists only of the unit element $\hat{E}$. Hence, every group G contains two trivial subgroups: $H = E$ and $H = G$ which are groups of the wave vector **k** at the common point and at the point Γ, respectively.

There are different notations for the symmetry points of various Bravais lattices.

1. Simple cubic lattice. The Brillouin zone of a simple cubic lattice is a cube with the edge $\frac{2\pi}{a}$ long (Fig. 145). Inner symmetry elements include the points:

(a) Δ—of the four-fold axis. There are altogether three such equivalent axes: k_x, k_y, k_z.

(b) Λ—of the three-fold axis (four equivalent axes).

(c) Σ—of the three-fold axis (six equivalent axes).

(d) of the symmetry planes ΔΣ, ΣΛ and ΛΔ.

Outer symmetry points include the points of the boundary planes and of their intersections. There are special notations for points which are both inner and outer symmetry points. They include the points X, Z, M, T, R and S. Their origin may be seen from Fig. 145.

The irreducible representations of the wave vector group are designated by the same letters as the symmetry points. For example, the irreducible representation of the wave vector group of the Δ-axis is designated by $\Delta^{(\alpha)}$. Find the group Δ. It follows directly from Fig. 145 that the elements of the group Δ include: E, rotations C_4, C_4^2, C_4^3, and reflections in two co-ordinate planes of the type ΔΣ and in two diagonal planes of the type ΛΔ. For the sake of simplicity denote these planes by σ_y, σ_z and σ_d' and σ_d''. We should make sure that the products of these elements do not form new symmetry elements. Obviously, the products of the planes result in already familiar rotations, for instance $\sigma_y \sigma_z = C_2$; $\sigma_d' \sigma_d'' = C_2$. The products of the type $\sigma_y \sigma_d'$ result in rotations C_4 or $C_4^3 = C_4^{-1}$. The direct consequence of this fact is that the aforementioned elements contain the products of rotations by reflections, as well. Indeed, the element of the type $C_4 \sigma_y$ may be represented in various forms, for example, $C_4 \sigma_y = C_4^{-1} C_4^2 \sigma_y = C_4^{-1} \sigma_z \sigma_y \sigma_y = C_4^{-1} \sigma_z = \sigma_d' \sigma_z \sigma_z = \sigma_d'$, etc. In other words, various combinations of the element products do not result in new elements, so the order of the group Δ is eight. Find the classes of the group. The four planes should be divided in two classes, $2\sigma_y$ and $2\sigma'_d$. The elements E, C_4 and C_2 constitute classes by themselves as elements of different order. It remains to be seen whether the element $C_4^3 = C_4^{-1}$ constitutes a separate class. Because of the existence of two pairs of reflection planes it is easy to prove that C_4^{-1} is conjugate with C_4. Indeed, it follows from $C_2 = \sigma_z\sigma_y$ and $C_2 = \sigma'_d\sigma''_d$ that $\sigma'_d = C_2\sigma''_d = \sigma''_d C_2 = \sigma''_d\sigma_y\sigma_z = C_4\sigma_z$, whence $\sigma_d'^{-1} = \sigma'_d = \sigma_z C_4^{-1} = C_4\sigma_z$ and $C_4 = \sigma_z C_4^{-1}\sigma_z^{-1}$. Hence, the group Δ has five classes: E, $2C_4$, C_2, σ_y, σ'_d, and five irreducible representations whose dimensionalities satisfy the equation

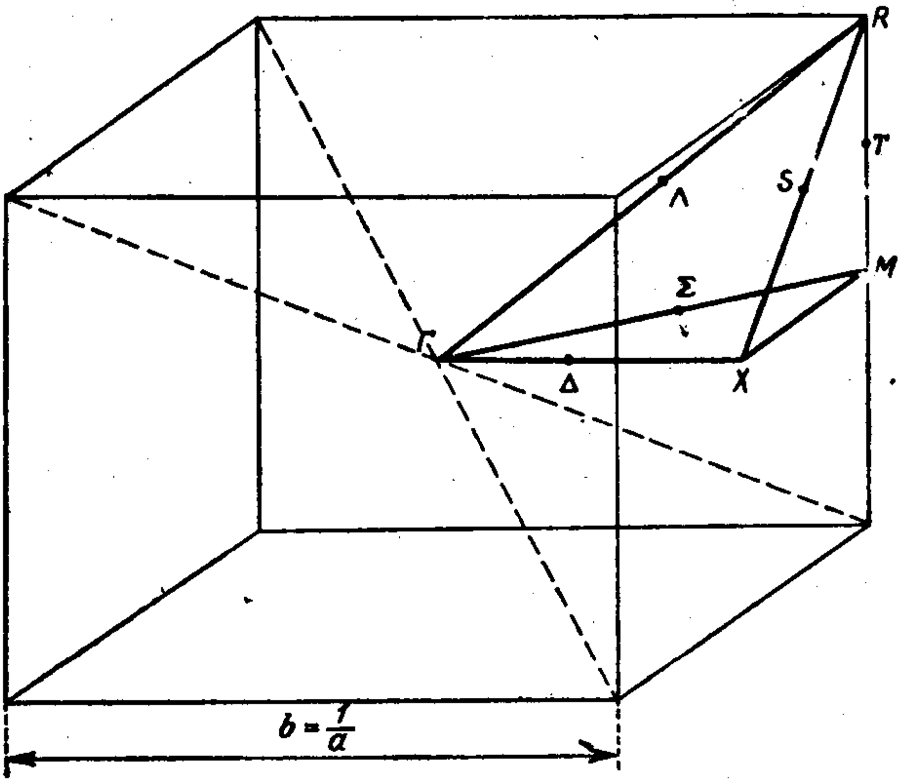


Fig. 145. Brillouin zone for a simple cubic lattice

$$1^2 + 1^2 + 1^2 + 1^2 + 2^2 = 8. \qquad (A.158)$$

The group Δ is isomorphous with the group $4mm$, or C_{4v}. As a result the table of characters of the group Δ coincides with that of the group C_{4v}. But this means that there is a substantial difference between the irreducible representations of the cubic full symmetry group O_h at the common point and in the states Δ. Since the group Δ is a subgroup of the group O_h, it is convenient to use the same notation for its classes as for those of group O_h. Taking into account that an arbitrary plane may be represented as a product of inversion by the two-fold rotation axis orthogonal to the inversion plane, we may write $\sigma = iC_2$. For this reason the notation $2iC_2$ and $2iC_4^2$ is used instead of $2\sigma_y$ and $2\sigma'_d$. For a specified axis Δ such a notation is symbolic because the multipliers are not themselves real elements of the group.

Below we present the characters of minor representations of Δ and of its isomorphous group at points T.

Characters of minor representations of Δ and *T*

Δ, T	E	C_4^2	$2C_4$	$2iC_4^2$	$2iC_2$
Δ_1	1	1	1	1	1
Δ_2	1	1	−1	1	−1
Δ_2'	1	1	−1	−1	1
Δ_1'	1	1	1	−1	−1
Δ_5	2	−2	0	0	0

Consider now the group Λ. It contains elements E; C_3^2, C_3 and three planes σ of the type ΛΣ. The product of two planes results in rotations C_3 and C_3^2, and the product of the elements of the type $C_3\sigma$ again yields planes. Hence, the group Λ has six elements. The planes σ constitute a single class. The elements C_3 and C_3^{-1} are conjugate which is easier to prove with the aid of the relation $XC_3X^{-1}=XC_3X$. If, for instance, C_3 is produced by the planes σ′ and σ″, then $C_3=\sigma'\sigma''$, whence $\sigma''\sigma'=C_3^{-1}$. Putting $X=\sigma'$ we will obtain $\sigma'C_3\sigma'=\sigma'\sigma'\sigma''\sigma'=\sigma''\sigma'=C_3^{-1}$. It follows that the group Λ has six elements and three classes: E, $2C_3$, 3σ. The dimensionalities of the irreducible representations are

$$1^2+1^2+2^2=6. \qquad (A.159)$$

The group Λ is isomorphous with the group $3m$, or C_{3v}.

Characters of minor representations of groups Λ and *F*

Λ, F	E	$2C_3$	$3iC_2$
Λ_1	1	1	1
Λ_2	1	1	−1
Λ_3	2	−1	0

In this case, too, the symmetry planes are designated symbolically as products of inversion by two-fold rotation axis to facilitate comparison with the O_h group.

As has already been stated, the irreducible representations of a group become reducible representations of a subgroup. The expansion of irreducible representations of the group $\Gamma^{(\alpha)}(\hat{G}_l)$ in irreducible representations of the wave vector group, i.e. in the subgroup of group G, is performed with the aid of usual rules for expanding of reducible representations based on the character orthogonality theorems. Write

$$\Gamma^{(\alpha)}(\hat{G}_l)=\sum_{\beta} a_{\beta}(\Gamma^{(\alpha)})\,\Delta_{\beta}.$$

The expansion coefficients may be found from the relation

$$a_{\beta}(\Gamma^{(\alpha)})=\frac{1}{h}\sum_{H_l} X_{\Gamma}^{(\alpha)}(H_l)\,X^{(\beta)*}(H_l), \tag{A.160}$$

where h is the order of the wave vector group, eight for the group Δ. When determining the expansion coefficients it is convenient to write out the irreducible representations of the wave vector group, for instance of Δ, and below them—the characters of the same classes in the $\Gamma^{(\alpha)}$ representations. Using the tables of characters of groups O_h and Δ write:

Table of characters of the group O_h

Notation											
T_d	O	1/E, W, K	2/B, S, W	3/L, B	4/H, J	E	$6C_4$	$6C_2$	$3C_4^2$	$3C_2$	$6iC_4$
$(\Gamma_1)_g$	$(\Gamma_1)_g$	A_{1g}	Γ_1	α	Γ_s	1	1	1	1	1	1
$(\Gamma_1)_u$	$(\Gamma_2)_u$	A_{2u}	Γ_2'	β	Γ_f'	1	−1	−1	1	1	1
$(\Gamma_2)_g$	$(\Gamma_2)_g$	A_{2g}	Γ_2	β'	Γ_f	1	−1	−1	1	1	−1
$(\Gamma_2)_u$	$(\Gamma_1)_u$	A_{1u}	Γ_1'	α'	Γ	1	1	1	1	1	−1
$(\Gamma_3)_g$	$(\Gamma_3)_g$	E_g	Γ_{12}	γ	Γ_d'	2	0	0	2	−1	0
$(\Gamma_3)_u$	$(\Gamma_3)_u$	E_u	Γ_{12}'	γ'	Γ_h	2	0	0	2	−1	0
$(\Gamma_4)_g$	$(\Gamma_4)_g$	T_{2g}	Γ_{25}'	ε	Γ_d^2	3	−1	1	−1	0	−1
$(\Gamma_4)_u$	$(\Gamma_4)_u$	T_{1u}	Γ_{15}	δ	Γ_p	3	1	−1	−1	0	−1
$(\Gamma_5)_g$	$(\Gamma_5)_g$	T_{1g}	Γ_{15}'	δ'	Γ_g	3	1	−1	−1	0	1
$(\Gamma_5)_u$	$(\Gamma_5)_u$	T_{2u}	Γ_{25}	ε'	Γ_f^2	3	−1	1	−1	0	1

1) E, W, K—Eyring, Walter, Kimball
2) B, S, W—Bowker, Smoluchowski, Wigner
3) L, B—Lage, Bethe
4) H, J—Howars, Jones

Δ	E	C_4^2	$2C_4$	$2iC_4^2$	$2iC_2$
Δ_1	1	1	1	1	1
Δ_2	1	1	-1	1	-1
Δ_2'	1	1	-1	-1	1
Δ_1'	1	1	1	-1	-1
Δ_5	2	-2	0	0	0
Γ_1	1	1	1	1	1
Γ_2	1	1	-1	1	-1
Γ_{12}	2	2	0	2	0
Γ_{15}'	3	-1	1	-1	-1
Γ_{25}'	3	-1	-1	-1	1
Γ_1'	1	1	1	-1	-1
Γ_2'	1	1	-1	-1	1
Γ_{12}	2	2	0	-2	0
Γ_{15}	3	-1	1	1	1
Γ_{25}	3	-1	-1	1	-1

Expanding in compliance with the formula (A.154) we obtain

$$\begin{array}{ll}
a_1(\Gamma_1)=1 & a_1'(\Gamma_1')=1 \\
a_2(\Gamma_2)=1 & a_2'(\Gamma_2')=1 \\
a_1(\Gamma_{12})=1 & a_1'(\Gamma_{12}')=1 \\
a_2(\Gamma_{12})=1 & a_2'(\Gamma_{12}')=1 \\
a_1'(\Gamma_{15}')=1 & a_1(\Gamma_{15})=1 \\
a_5(\Gamma_{15}')=1 & a_5(\Gamma_{15})=1 \\
a_2(\Gamma_{25}')=1 & a_2(\Gamma_{25})=1 \\
a_5(\Gamma_{25}')=1 & a_5(\Gamma_{25})=1,
\end{array} \qquad \text{(A.161)}$$

other coefficients being zero.

The representations $\Gamma^{(\alpha)}$ are seen to be reduced in the subgroup Δ. Of maximum interest is the reduction of multidimensional representations.

Two-dimensional representations Γ_{12} and Γ_{12}' are seen to split into two unidimensional representations Δ_1 and Δ_2 and Δ_1' and Δ_2', respectively. Three-dimensional representations Γ_{25} and Γ_{25}' also split into two representations: a unidimensional Δ_2 and Δ_2' and a two-dimensional Δ_5. The expansion of irreducible representations of point groups in irreducible representations of the wave vector group, i.e. the reduction on a subgroup, is known in the solid-state theory by the name of compatibility relation determination. These relations determine the nature of the variations of the solid's physical properties resulting from the transformation of the common points of the Brillouin zone into the symmetry points.

The tables of compatibility have been compiled for the majority of the more important crystals. We present below, by way of example, the tables of compatibilfty for some symmetry points.

Compatibility relations between Γ and Δ, Λ, and Σ

Γ	Δ	Λ	Σ
Γ_1	Δ_1	Λ_1	Σ_1
Γ_2	Δ_2	Λ_2	Σ_2
Γ_{12}	$\Delta_1\Delta_2$	Λ_3	$\Sigma_1\Sigma_4$
Γ'_{15}	$\Delta'_1\Delta_5$	$\Lambda_2\Lambda_3$	$\Sigma_2\Sigma_3\Sigma_4$
Γ'_{25}	$\Delta'_2\Delta_5$	$\Lambda_1\Lambda_3$	$\Sigma_1\Sigma_2\Sigma_3$
Γ'_1	Δ'_1	Λ_2	Σ_2
Γ'_2	Δ'_2	Λ_1	Σ_3
Γ'_{12}	$\Delta'_1\Delta'_2$	Λ_3	$\Sigma_2\Sigma_3$
Γ_{15}	$\Delta_1\Delta_5$	$\Lambda_1\Lambda_3$	$\Sigma_1\Sigma_3\Sigma_4$
Γ_{25}	$\Delta_2\Delta_5$	$\Lambda_2\Lambda_3$	$\Sigma_1\Sigma_2\Sigma_4$

Note the following property of the compatibility tables. The representations Γ_{25} and Γ'_{25} at the common poins of the representations are three-dimensional. On the Δ-axis they split into two representations, one of them unidimensional and the other two-dimensional. But on the Σ-axis Γ_{25} and Γ'_{25} split into three unidimensional representations: $\Sigma_1+\Sigma_2+\Sigma_4$ and $\Sigma_1+\Sigma_2+\Sigma_3$, respectively. This is very important for solving the problem of degeneration multiplicity of the energy levels in the Brillouin zone.

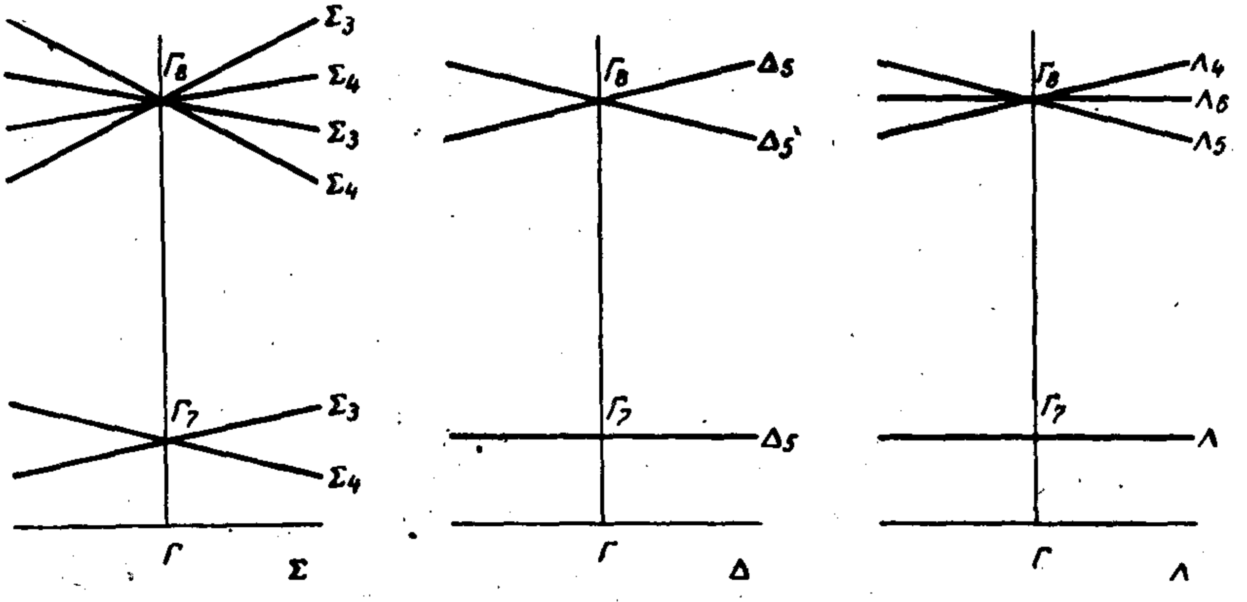


Fig. 146. Splitting of the Γ_7 and Γ_8 levels along the axes Σ, Δ and Λ in a double group

2. **Face-centered lattice.** The points of symmetry of the Brillouin zone are shown in Fig. 146. As in the simple cubic lattice the elements of distinction are the four-, three- and two-fold axes Δ, Λ and Σ. Minor representations and compatibility relations for the points of these axes are analogous to those of the simple cubic lattice.

The characters of the irreducible representations of Σ

Σ	E	C_2	iC_4^2	iC_2
Σ_1	1	1	1	1
Σ_2	1	1	-1	-1
Σ_3	1	-1	-1	1
Σ_4	1	-1	1	-1

As these axes transgress the boundary of the Brillouin zone, the points X, L and K appear.

(a) Point L. The order of the wave vector group at point L is twelve. The distribution of the elements among the classes is as follows:

$$\begin{aligned} L&:E;\ 2C_3=(C_3,\ C_3^{-1});\quad 3iC_2=(m_1,\ m_2,\ m_3);\\ i&:2iC_3=(iC_3,\ iC_3^{-1});\quad 3C_2=(im_1,\ im_2,\ im_3). \end{aligned} \tag{A.162}$$

The appearance of the element i and of the five elements $2iC_3$ and $3C_2$ connected with it is due to the fact that the definition of the group takes it for granted that $G_k\mathbf{k}=\mathbf{k}+2\pi\mathbf{b}$, i.e. that the invariance of $\mathbf{k}$ includes coincidence of L with an equivalent point which may be obtained by the addition of the appropriate vector $2\pi\mathbf{b}$, this being equivalent to inversion. Since the group L may be regarded as equal to $\Lambda\times C_i$, the number of elements, classes and irreducible representations is doubled. The table of characters may be obtained from the tables for Λ and C_i.

There are altogether four points of the type L, the co-ordinates of one of them being $(\pi/a,\ \pi/a,\ \pi/a)$, where a is the lattice parameter.

(b) Point X. There are in all three points of the type X with coordinates of the type $(2\pi/a,\ 0,\ 0)$. The group X contains, in addition to the axes and planes of symmetry, also the inversion, this being due to the fact that the points $\mathbf{x}$ and $\mathbf{x}+2\pi\mathbf{b}$ are equivalent. The order of the group X is 16, and the number of classes 10. This results in eight unidimensional and two two-dimensional

The characters of minor representations of group L

	Λ			$i\Lambda$			Basic functions
	E	$2C_3$	$3iC_2$	i	$2iC_3$	$3C_2$	
L_1	1	1	1	1	1	1	1
L_2	1	1	−1	1	1	−1	$xy(x^2-y^2)+yz(y^2-z^2)+$ $+zx(z^2-x^2-y^2)$
L_3	2	−1	0	2	−1	0	$z^2-\frac{1}{2}(x^2+y^2)$
L_1'	1	1	1	−1	−1	−1	$(x-y)(y-z)(z-x)$
L_2'	1	1	−1	−1	−1	1	$x+y+z$
L_3'	2	−1	0	−2	1	0	$x-z;\ y-z$

representations. The group X is isomorphous with the tetragonal group $4/m$ mm.

The examples presented above make the principle of construction of minor representations and the method of finding the compatibility relations sufficiently clear.

11. SCHRÖDINGER EQUATION

One may find the properties of the crystal by solving the Schrödinger equation

$$\hat{H}\psi = E\psi \tag{A.163}$$

determined by the Hamiltonian $\hat{H}$. In the single-electron approximation the Hamiltonian $\hat{H}$ may be represented as the sum of the kinetic $\hat{T} = -(\hbar^2/2m)\Delta$ and the potential $\hat{U}(\mathbf{r})$ energy operators. Since $U(\mathbf{r})$ is the potential energy of the electron in the crystal we are in a position to assert that all transformations which result in the lattice coinciding with itself should leave the value of the potential energy at the selected point $M(\mathbf{r})$ unchanged. In other words, we may assert that the potential energy is an invariant with respect to the transformations of symmetry of the lattice space group:

$$\hat{G}_i U(\mathbf{r}) = U(\hat{G}_i \mathbf{r}) = U(\mathbf{r}). \tag{A.164}$$

Should we regard $\hat{U}(\mathbf{r})$ as an operator identical with the operator of multiplication by $U(\mathbf{r})$ we would be able to identify the condition of invariance of $U(\mathbf{r})$ as the condition of commutation of the operators

$$\hat{G}_i\hat{U}(\mathbf{r}) = \hat{U}(\mathbf{r})\hat{G}_i \qquad \text{(A.165)}$$

which can be easily proved:

$$\hat{G}_i\hat{U}(\mathbf{r})\,\psi(\mathbf{r}) = \hat{G}_i U(\mathbf{r})\,\psi(\mathbf{r}) = U(\hat{G}_i\mathbf{r})\,\psi(\hat{G}_i\mathbf{r}) =$$
$$= U(\mathbf{r})\,\hat{G}_i\psi(\mathbf{r}) = \hat{U}(\mathbf{r})\,\hat{G}_i\psi(\mathbf{r}),$$

where ψ is an arbitrary function.

The kinetic energy operator $\hat{T}$ remains invariant in all transformations belonging to the space symmetry group.

This assertion is quite obvious as regards translation.

It is easily proved that for the rotation $\mathbf{r}' = \hat{R}\mathbf{r}$ with a matrix A for which $\det A = 1$, $\Delta' = \Delta$.

An improper rotation is determined by the matrix A for which $\det A = -1$, the orthogonality condition remaining unchanged.

Hence, the result is that the Hamiltonian commutes with all the operators $\hat{G}_i$ of the crystal symmetry group G:

$$\hat{G}_i\hat{H} = \hat{H}\hat{G}_i. \qquad \text{(A.166)}$$

This, in turn, means that the eigenfunctions of the Hamiltonian are eigenfunctions of the operators of the group G. Thus, we are able to obtain substantial information on the wave functions $\psi(\mathbf{r})$ and the eigenvalues E of the Hamiltonian $\hat{H}$ without solving the Schrödinger equation and even without writing it out in "explicit" form. For instance, the eigenfunctions of the Hamiltonian should be of the form of the Bloch wave:

$$\psi(\mathbf{r}) = \psi_{\mathbf{k}}(\mathbf{r}) = e^{i(\mathbf{kr})}\,\psi_{\mathbf{k}}(\mathbf{r}), \qquad \text{(A.167)}$$

where $\mathbf{k}$ is the wave vector which determines the irreducible representations of the translational group, and $\psi_{\mathbf{k}}(\mathbf{r})$ should be a periodic function: $\psi_{\mathbf{k}}(\mathbf{r}) = \psi_{\mathbf{k}}(\mathbf{r} + \mathbf{n})$.

We are able to forecast the behaviour of $\psi_{\mathbf{k}}(\mathbf{r})$ in the course of inversion, reflection in planes and other operations of symmetry where it would turn zero (nodal planes). The Bloch functions for an electron in a crystal of a specified symmetry should possess all the properties of the basic functions of appropriate symmetry contained in the tables of characters.

Moreover, even without substituting $\psi_{\mathbf{k}}(\mathbf{r})$ into the equation we may assert that the energy E is a function of $\mathbf{k}$ since definite eigenvalues $E = E(\mathbf{k})$ correspond to the eigenfunctions $\psi_{\mathbf{k}}(\mathbf{r})$. The irreducible representations will be equivalent if $\mathbf{k}' = \mathbf{k} + 2\pi\mathbf{b}$. But this means that energy values $E(\mathbf{k}') = E(\mathbf{k})$ since $E(\mathbf{k})$ should be independent of the concrete form of the basic functions connected with the irreducible representations.

The properties of the $E(\mathbf{k})$ functions are also for the most part known to us because they are determined by the point group of the crystal symmetry expressed in terms of the Brillouin zone symmetry.

Let us discuss the problem at some length.

Let the solution of the equation

$$\hat{H}\psi_{\mathbf{k}} = E\psi_{\mathbf{k}} \tag{A.168}$$

be f-fold degenerate, i.e. there are f different wave functions $\psi_{\mathbf{k}1}$, $\psi_{\mathbf{k}2}, \ldots, \psi_{\mathbf{k}f}$ corresponding to a definite value of $E=E(\mathbf{k})$. We will assume them to be orthonormalized:

$$\int \psi_{\mathbf{k}\alpha}^{*}\psi_{\mathbf{k}\beta}\,d\tau = \delta_{\alpha\beta}. \tag{A.169}$$

Apply the operators of the group G to the equation

$$\hat{H}\psi_{\mathbf{k}\alpha} = E(\mathbf{k})\,\psi_{\mathbf{k}\alpha}, \quad \alpha = 1\ \ 2, \ldots, f. \tag{A.170}$$

If translation operators are applied, the result will be the one already familiar to us—$\psi_{\mathbf{k}\alpha}$ will vary according to irreducible representations of the translational group $e^{i(\mathbf{kn})}$ which is brought about by the choice of $\psi_{\mathbf{k}\alpha}$ in the form of the Bloch wave. Apply now the operators of the point group G/T or, in short, simply G:

$$\hat{G}_l\hat{H}\psi_{\mathbf{k}\alpha} = \hat{H}\hat{G}_l\psi_{\mathbf{k}\alpha} = E(\mathbf{k})\,\hat{G}_l\psi_{\mathbf{k}\alpha}. \tag{A.171}$$

We see from here that if $\psi_{\mathbf{k}\alpha}$ are the solutions of the equations they may be used to obtain new solutions of the type $\hat{G}_l\psi_{\mathbf{k}\alpha}$. But since the dimensionality of the space has already been fixed it follows that the new solutions cannot be linearly independent and may therefore be expanded in the basis. The coefficients of expansion are already known to us—they are the matrix elements of the group G calculated with the basic functions $\psi_{\mathbf{k}\alpha}$:

$$\Gamma_{\alpha\beta}(\hat{G}_l) = \int \psi_{\mathbf{k}\alpha}^{*}\hat{G}_l\psi_{\mathbf{k}\beta}\,d\tau. \tag{A.172}$$

Hence, we may say that the eigenfunctions $\psi_{\mathbf{k}\alpha}$ are transformed with the aid of the representation $\Gamma(\hat{G}_l)$ of the group G:

$$\hat{G}_l\begin{pmatrix}\psi_{\mathbf{k}1}\\ \vdots \\ \psi_{\mathbf{k}f}\end{pmatrix} = \tilde{\Gamma}(\hat{G}_l)\begin{pmatrix}\psi_{\mathbf{k}1}\\ \vdots \\ \psi_{\mathbf{k}f}\end{pmatrix}. \tag{A.173}$$

Two ca es are possible:

1. $\Gamma(\hat{G}_l)$ is irreducible: the eigenfunctions of a degenerate state are transformed with the aid of one of the irreducible representations of the group G. The degeneracy is termed natural, or essential.

2. $\Gamma(\hat{G}_l)$ is reducible: $\Gamma(\hat{G}_l)=\sum_{\alpha} a_\alpha \Gamma^{(\alpha)}(\hat{G}_l)$. The eigenfunctions may be expanded in invariant subspaces which are transformed with the aid of appropriate irreducible representations. The state degeneracy in this case is termed fortuitous. The meaning of the term will be made clear below.

As we have seen from the above, the greatest dimensionality of irreducible representations of the point groups is 3. In consequence, essential degeneracy due, specifically, to the symmetry of the state cannot exceed 3.

Apply a perturbation $\hat{W}$ to the quantum system. Two cases of principal interest are possible:

1) The symmetry group of $\hat{W}$ is higher or coincides with the symmetry group of the lattice field $\hat{U}(\mathbf{r})$. The degeneracy may be removed by an appropriate choice of $\hat{W}$. The term fortuitous is applied just to the degeneracy occasioned by the magnitude of the field $W(\mathbf{r})$, or $U(\mathbf{r})$. However, the eigenfunctions which are transformed with the aid of different irreducible representations may not belong to different eigenvalues since this would be tantamount to the destruction of the irreducible representation. The irreducibility of the representation is, however, determined not by the magnitude, but by the symmetry of the field. In other words, the degeneracy occasioned by the symmetry may not be removed by the application of a perturbation.

2) The symmetry of the perturbation is of the lower order than that of the lattice field $U(\mathbf{r})$. The full symmetry of the Hamiltonian will i. this case be determined by the "intersection" of the crystal and the perturbation symmetry groups, i.e. by the elements common to both.

Suppose there is no fortuitous degeneracy. The question arises: will the natural degeneracy be removed? The answer is in the compatibility relations: the irreducible representations of the group of the unperturbed system should be expanded in the irreducible representations of the perturbed system with the result that the states whose representations will now become reducible will be decomposed into states of lesser degeneracy, or into $f^{(\alpha)}$ non-degenerate states, in case the system has unidimensional representations corresponding to the original degenerate state.

Thus, the theory of groups enables us to find the degree of degeneracy of the states and the way in which degeneracy is removed by perturbation by resorting solely to the properties of symmetry.

We considered the behaviour of the wave functions when the symmetry transformation was applied to the Schrödinger equation. Consider now the eigenvalue $E(\mathbf{k})$. The vector $\mathbf{k}$ was introduced as a quantity which determines the representation of the translational group $e^{i(\mathbf{kn})}$. By performing some space transformation $\hat{G}_l$

we will transform the vector $\mathbf{n}$, as well: $\mathbf{n}' = \hat{G}_l\mathbf{n}$ with the result that the representation will change by $e^{i(\mathbf{kn}')}$. If we should now apply the same transformation $\hat{G}_l$ to the $\mathbf{k}$-space the result would be $\mathbf{k}' = \hat{G}_l\mathbf{k}$, so that $(\mathbf{k}'\mathbf{n}') = (\mathbf{kn})$.

We take account of the fact that the vectors $\mathbf{k}$ and $\mathbf{n}$ are real, and that $\hat{G}_l$ may be reduced to orthogonal transformations, so that formally we obtain

$$(\hat{G}_l\mathbf{k}, \hat{G}_l\mathbf{n}) = (\mathbf{k}, \hat{\hat{G}}_l\hat{G}_l\mathbf{n}) = (\mathbf{k}, \hat{G}_l^{-1}\hat{G}_l\mathbf{n}) = (\mathbf{kn}). \qquad (A.174)$$

We made use of the condition of invariance of the scalar product of real vectors in orthogonal transformations. But this means that for irreducible representations, too, $e^{i(\mathbf{kn})} = e^{i(\mathbf{k}'\mathbf{n}')}$ with the result that the eigenfunctions $\psi_{\mathbf{k}}(\mathbf{r})$ and, consequently, the eigenvalues remain invariant. Hence, we arrive at the conclusion that

$$\hat{G}_lE(\mathbf{k}) = E(\hat{G}_l\mathbf{k}) = E(\mathbf{k}), \qquad (A.175)$$

i.e. that the symmetry group of energy in the Brillouin zone is the crystal symmetry point group.

12. TWIN GROUPS. TIME INVERSION

When we considered the motion of the electron in a crystal we took no account of the spin. Consider now the effect of the spin. As was demonstrated in Sec. A2 the energy of the interaction of the electron magnetic moment $\boldsymbol{\mu} = \frac{e}{m_0c}\mathbf{S}$ with the electric field of the lattice $\mathbf{E} = -\frac{1}{e}\nabla U(\mathbf{r})$ may be presented in the form

$$W = \frac{1}{2m^2c^2}(\hat{\mathbf{S}}[\nabla U\hat{\mathbf{p}}]). \qquad (A.176)$$

Denoting the Hamiltonian without the spin by $\hat{H}_0$,

$$\hat{H}_0 = \frac{\hat{\mathbf{p}}^2}{2m}\Delta + U(\mathbf{r}), \qquad (A.177)$$

we write

$$\hat{H} = \hat{H}_0 + \hat{W}. \qquad (A.178)$$

To investigate the group properties of H reduce it to a form analogous to that used in the Pauli equation. In operator form the spin

$\hat{\mathbf{S}}$ is determined by three Pauli matrices $\hat{\sigma}_x$, $\hat{\sigma}_y$, $\hat{\sigma}_z$:

$$\begin{aligned}\hat{S}_x&=\frac{\hbar}{2}\hat{\sigma}_x=\frac{\hbar}{2}\begin{pmatrix}0&1\\1&0\end{pmatrix};\\ \hat{S}_y&=\frac{\hbar}{2}\hat{\sigma}_y=\frac{\hbar}{2}\begin{pmatrix}0&-i\\i&0\end{pmatrix};\\ \hat{S}_z&=\frac{\hbar}{2}\hat{\sigma}_z=\frac{\hbar}{2}\begin{pmatrix}1&0\\0&-1\end{pmatrix};\end{aligned}\tag{A.179}$$

The Pauli matrices have the following properties:

$$\begin{gathered}\hat{\sigma}_l^2=1;\quad r=x,\,y,\,z;\\ 1=\begin{pmatrix}1&0\\0&1\end{pmatrix}.\end{gathered}\tag{A.180}$$

Putting $\hat{\mathbf{S}}=\frac{\hbar}{2}\hat{\sigma}$ re-write $\hat{\mathbf{W}}$

$$\hat{\mathbf{W}}=\frac{\hbar}{4m^2c^2}([\nabla U,\hat{\mathbf{p}}]\,\hat{\sigma})=-\frac{i\hbar^2}{4m^2c^2}([\nabla U,\,\nabla]\,\sigma).\tag{A.181}$$

Introduce the vector operator $\hat{\mathbf{N}}$:

$$\hat{\mathbf{N}}=\left(\frac{\hbar}{2mc}\right)^2[\nabla U,\,|\nabla].\tag{A.182}$$

Its components are of the form

$$\begin{aligned}N_x=N_1&=\left(\frac{\hbar}{2mc}\right)^2\left(\frac{\partial u}{\partial y}\frac{\partial}{\partial z}-\frac{\partial u}{\partial z}\frac{\partial}{\partial y}\right)=N_{23};\\ N_y=N_2&=\left(\frac{\hbar}{2mc}\right)^2\left(\frac{\partial u}{\partial z}\frac{\partial}{\partial x}-\frac{\partial u}{\partial x}\frac{\partial}{\partial z}\right)=N_{31};\\ N_z=N_3&=\left(\frac{\hbar}{2mc}\right)^2\left(\frac{\partial u}{\partial x}\frac{\partial}{\partial y}-\frac{\partial u}{\partial y}\frac{\partial}{\partial x}\right)=N_{12}.\end{aligned}\tag{A.183}$$

One may write in a general form

$$N_r=\frac{1}{2}\sum_{p,\,q}\varepsilon_{rpq}N_{pq},\tag{A.184}$$

where ε_{rpq} is an element of the unit antisymmetrical tensor which is equal to zero if at least two of its indices are identical. The other elements of the tensor are equal to ±1, so that

$$\varepsilon_{123}=\varepsilon_{231}=\varepsilon_{312}=1\tag{A.185}$$

and

$$\varepsilon_{213}=\varepsilon_{132}=\varepsilon_{321}=1.\tag{A.186}$$

The element of the tensor may be expressed in terms of the vector component as well:

$$N_{pq}=\sum_r N_r e_{rpq}.$$

The vectors ∇U and ∇ are polar vectors. The vector product of two polar vectors is an axial vector, and for this reason the indices of some components of **N** are those of vector components (one index), and of others—those of an antisymmetrical tensor (two indices).

Making use of $\hat{\mathrm{N}}$ and $\hat{\sigma}$ we may write $\hat{\mathrm{W}}$ as follows:

$$\hat{\mathrm{W}}=-i(\hat{\mathrm{N}}\hat{\sigma})=-i\sum_{k=1}^{3}\hat{\mathrm{N}}_k\hat{\sigma}_k. \qquad (A.187)$$

The existence of the two-row matrix $\hat{\sigma}$ determines the choice of the wave function in the form of a two-row column:

$$\Psi=\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}. \qquad (A.188)$$

The Schrödinger equation may be written in the form

$$\hat{\mathrm{H}}\Psi=\{\hat{\mathrm{H}}_0-i(\hat{\mathrm{N}}\hat{\sigma})\}\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}=E\begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix}=E\Psi. \qquad (A.189)$$

Let G be a point group of the order g of the Hamiltonian $\hat{\mathrm{H}}_0$. Find the group of the Hamiltonian $\hat{\mathrm{H}}$. Take some operator $\hat{\mathrm{G}}_j$ with the corresponding matrix A so that the transformation

$$\mathbf{r}'=\hat{\mathrm{G}}_j\mathbf{r} \qquad (A.190)$$

may be written in the form

$$x'_k=\sum_l a_{kl}x_l. \qquad (A.191)$$

Recall that for a proper rotation det $A=1$, and for an improper rotation det $A=-1$, the rows (and columns) being orthonormalized.

The vector operator $\hat{\mathrm{N}}$ may be regarded either as an antisymmetrical tensor of rank II, or as a vector dual to it. It is established in vector analysis that in this case the vector is axial. The difference between a polar and an axial vector stands out when the inversion of the co-ordinate axes (or of space) is performed: a polar vector changes sign, the axial does not. Making direct calculations or applying tensor component transformation rules obtain that

$$N'_{pq}=\sum_{l,\,m}a_{pl}a_{qm}N_{lm}. \qquad (A.192)$$

One may make use of the properties of the antisymmetrical tensor ε_{rst} and find the vector components N_j transformation law. For example, putting $p=2$, $q=3$ find that

$$N'_1 = N'_{23} = A''N_1 + A^{(12)}N_2 + A^{(13)}N_3 = \sum_{l=1}^{3} A^{(1l)}N_l, \quad \text{(A.193)}$$

where $A^{(1l)}$ is the algebraic supplement to the element a_{1l} of the matrix A. We may write, from the condition

$$\sum_{l=1}^{3} a_{1l}A^{(1l)} = \pm 1 \quad \text{(A.194)}$$

that

$$N'_1 = \pm \sum_{l=1}^{3} a_{1l}N_l \quad \text{(A 195)}$$

or, generally,

$$N'_j = \pm \sum_{l=1}^{3} a_{jl}N_l. \quad \text{(A.196)}$$

The sign (+) refers to a proper rotation, (—)—to an improper rotation or inversion.

Leaving out the signs write the reciprocal transformation which, owing to the orthogonality of the matrix A, will be determined by the transposed matrix $A^{-1} = \tilde{A}$:

$$N_j = \sum_{k=1}^{3} a_{kj}N_k. \quad \text{(A.197)}$$

Write the Hamiltonian in the form

$$\hat{H} = \hat{H}_0 - i\sum_k \hat{N}_k\hat{\sigma}_k. \quad \text{(A.198)}$$

Apply to $\hat{H}$ the operator $\hat{G}_j$ of the point group. When applying the operator $\hat{G}_j$ we should keep in mind that the co-ordinate axes remain unchanged with transformations of space; the spin of the particle, too, remains unchanged, therefore $\hat{G}_j$ does not affect $\hat{\sigma}_k$, and we obtain

$$\hat{G}_j\hat{H}(\mathbf{r}) = \hat{G}_j\hat{H}_0 - i\sum_k \hat{G}_j\hat{N}_k\hat{\sigma}_k$$

$$\hat{H}(\mathbf{r}') = \hat{H}_0(\mathbf{r}') - i\sum_k\sum_l a_{kl}\hat{N}_l\hat{\sigma}_k = \hat{H}_0(\mathbf{r}') - i\sum_l \hat{N}_l\Big(\sum_k a_{kl}\hat{\sigma}_k\Big). \quad \text{(A.199)}$$

Since $\hat{G}_j$ is an element of the crystal-symmetry point group it follows that $\hat{H}_0(\mathbf{r}') = \hat{H}_0(\mathbf{r})$. For the spin part of the Hamiltonian to remain invariant one should be able to substitute σ_l for the term $\sum_h a_{k\,l}\sigma_k$. Generally, this is impossible. Suppose that there exists a unitary matrix S which does not change the form of $\hat{H}_0$ and $\hat{N}$, but which transforms the co-ordinates of the spin

$$S^{-1}\left(\sum_k a_{kl}\sigma_k\right)S=\sigma'_l. \tag{A.200}$$

In this case we may write

$$S^{-1}\hat{H}(\hat{G}_j r)S=\hat{H}(r')=\hat{H}(r). \tag{A.201}$$

From the equation

$$\hat{H}\Psi=E\Psi;\quad \hat{G}_j\hat{H}\Psi=EG_j\Psi \tag{A.202}$$

one may obtain

$$S^{-1}\hat{H}\Psi S=ES^{-1}\Psi S \tag{A.203}$$

or

$$S^{-1}\hat{H}(r')SS^{-1}\hat{G}_j\Psi S=ES^{-1}\hat{G}_j\Psi S, \tag{A.204}$$

i.e. if Ψ is a solution of the Schrödinger equation so, too, will be $S^{-1}\hat{G}_j\Psi S$.

Investigations have shown that there are two matrices S_+ and S_- which correspond to every operator $\hat{G}_j$ and provide for the invariance of the Hamiltonian containing the spin term. When a rotation through the angle 2π is performed, $S_+\longrightarrow -S_-$.

Because of this the symmetry group $H(r)$ is made up of twice the number of elements and is termed double group. Bethe suggested that the rotation through the angle 4π be regarded as unit rotation, and the rotation through the angle 2π be denoted by $\mathbf{E}$. The corresponding matrix is $(-1)E$ where E is the unit matrix, so that $\mathbf{E}^2=E$.

The double group **G** may be represented as a direct product of the group G and $\{E,\ \mathbf{E}\}$:

$$\mathbf{G}=G\times\{E,\ \mathbf{E}\}. \tag{A.205}$$

The irreducible representations of a double group may be obtained in the form of the direct product of the irreducible representations of the groups G and $\{E,\ \mathbf{E}\}$. In other words, the irreducible representations of the double group contain all the irreducible representations of the single group G and of the group $G\times\mathbf{E}$. The number of classes in $G\times\mathbf{E}$ may not exceed the number of classes in $G\times E$; actually, it may be less. The element **E** constitutes its own class. The symmetry axis C_n becomes now a symmetry axis C_{2n}. Indeed, in the double group $(C_n)^n = \mathbf{E}$ and $(C_n)^{2n} = \mathbf{E}^2 = E$. The elements C_n^k and C_n^{2n-k} are mutually reciprocal: $C_n^{2n-k} = \mathbf{E} \times C_n^{n-k}$. For $n=2$ and $k=1$ we obtain $C_2^{2\times 2-1} = C_2^3 = \mathbf{E}C_2$. Since the axis C_2 is two-sided, $C_2^{-1} = C_2$, the rotations through the angles π and 3π belong to the same class, i.e. the elements C_2 and $\mathbf{E}C_2$ belong to the same class. Generally, the rotations about the axis C_n through the angles π and 3π will belong to one class if C_n is a two-sided axis, for instance, if it is perpendicular to the axis C_2 or σ. For reflections we obtain $\sigma^2 = \mathbf{E}$ and $\sigma^4 = \mathbf{E}^2 = E$, but $i^2 = E$, i.e. the elements $i\mathbf{E}$ and i belong to the same class.

By way of example, consider the double group of the point group T_d, or $43m$. The group T_d contains 24 elements distributed over 5 classes: E, $3C_2$, $8C_3$, $6S_4$, $6\sigma_d$; or in notations analogous to the group O_h: E, $3C_4^2$, $8C_3$, $6iC_4$, $6iC_2$. The double group $\mathbf{T}_d$ should contain the same elements and classes as the result of multiplication $T_d \times E$ and, in addition, 24 elements resulting from the multiplication $T_d \times \mathbf{E}$: $\mathbf{E}$, $3C_4^2\mathbf{E}$, $8C_3\mathbf{E}$, $6iC_4\mathbf{E}$ and $6iC_2\mathbf{E}$. The elements $\mathbf{E}$, $C_3\mathbf{E}$ and $iC_4\mathbf{E}$ should, undoubtedly, form new classes. The elements $C_4^2\mathbf{E}$ should join the class C_4^2 because the axis C_4^2 is two-sided. The elements $iC_2\mathbf{E}$ belong to the class iC_2 because iC_2 is a two-sided and a two-fold axis. Thus, the group $\mathbf{T}_d$ contains 48 elements and 8 classes. The three additional, termed spinor, irreducible representations are denoted, according to Bethe, by Γ_6, Γ_7

The characters of irreducible representations of the double group $\mathbf{T}_d$

	$\mathbf{T}_d$	E	$\mathbf{E}$	$6C_4^2$	$8C_3$	$8\mathbf{C}_3$	$6iC_4$	$6i\mathbf{C}_4$	$12iC_2$
$\Gamma_1\Gamma_1'$	Γ_1	1	1	1	1	1	1	1	1
$\Gamma_2\Gamma_2'$	Γ_2	1	1	1	1	1	−1	−1	−1
$\Gamma_{12}\Gamma_{12}'$	Γ_3	2	2	2	−1	−1	0	0	0
$\Gamma_{25}\Gamma_{25}'$	Γ_4	3	3	−1	0	0	−1	−1	1
$\Gamma_{15}\Gamma_{15}'$	Γ_5	3	3	−1	0	0	1	1	−1
	Γ_6	2	−2	0	1	−1	$\sqrt{2}$	$-\sqrt{2}$	0
	Γ_7	2	−2	0	1	−1	$-\sqrt{2}$	$\sqrt{2}$	0
	Γ_8	4	−4	0	−1	1	0	0	0
	Sp S^{-1}	2	−2	0	1	−1	$\sqrt{2}$	$-\sqrt{2}$	0

and Γ_8. Their dimensionalities may be found from the equation

$$48=24+24=24+2^2+2^2+4^2. \qquad (A.206)$$

The characters of the irreducible representations of the double group T_d are shown in the table.

The characters of the double group of representations Γ_6, Γ_7 and Γ_8 originate from the characters of the representations Γ_1, Γ_2 and Γ_3 of the single group through the action of the matrix S^{-1}. The matrix S may be written in the form:

$$S=\begin{pmatrix} \cos\frac{\beta}{2}e^{-\frac{i}{2}(\alpha+\gamma)} & -\sin\frac{\beta}{2}e^{-\frac{i}{2}(\alpha-\gamma)} \\ \sin\frac{\beta}{2}e^{\frac{i}{2}(\alpha-\gamma)} & \cos\frac{\beta}{2}e^{-\frac{i}{2}(\alpha+\beta)} \end{pmatrix} \qquad (A.207)$$

where α, β are Euler angles which determine the direction of the rotational axis, and γ is the angle of rotation about this axis.

It may be seen from the explicit expression for S that the increase in any angle of 2π transforms S into $-S$, this being equivalent to the introduction of the element $\bar{E}$. The increase in any angle of 4π results in $S\rightarrow S$.

Two-digit representations appear as $S^{-1}\Gamma^{(\alpha)}$. However, only $S^{-1}\Gamma_1$, $S^{-1}\Gamma_2$ and $S^{-1}\Gamma_3$ turn out to be irreducible. The representations $S^{-1}\Gamma_4$ and $S^{-1}\Gamma_5$ of dimensionality $f=6$ should, on the other hand, be reducible. The following relationship between $\Gamma^{(\alpha)}$ and $S^{-1}\Gamma^{(\alpha)}$ $(\alpha\leqslant 5)$ may be obtained:

$\Gamma^{(\alpha)}$	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5
$S^{-1}\Gamma^{(\alpha)}$	Γ_6	Γ_7	Γ_8	$\Gamma_7+\Gamma_8$	$\Gamma_6+\Gamma_8$

This simple table enables an important physical conclusion to be drawn. When the spin-orbital interaction is neglected the states Γ_1 and Γ_2 are singular, the state Γ_3 double degenerate, and the states Γ_4 and Γ_5 triple degenerate. Should the spin be taken into account, the degeneracy of each state would be doubled, i.e. the states Γ_1 and Γ_2 would become double degenerate, Γ_3—four-fold and Γ_4 and Γ_5—six-fold degenerate.

If, together with the spin, we take into account the spin-orbital interaction, the degeneracy of the states will be determined by the dimensionality of the irreducible representations of the double group. In this case the original states will remain unchanged, the states Γ_6 and Γ_7 will become double degenerate, and Γ_8—four-fold degenerate. The states Γ_4 and Γ_5 which now become six-fold degenerate will be decomposed into one four-fold degenerate and one double degenerate (Γ_6 or Γ_7) state. If now the spin degeneracy is once again neglected, the triple degenerate states Γ_4 and Γ_5 may be said to decompose, on account of spin-orbital interaction, into a double degenerate level Γ_8 and a single degenerate level Γ_7 or Γ_6, respectively. This is the usual method of formulating the result of the interaction of the spin with the lattice field.

To estimate the behaviour of the energy levels upon displacement from the point Γ of the Brillouin zone along the axes Δ, Λ and Σ it is necessary to find the double group of the wave vector and the compatibility relations. The results are shown in the tables.

The characters of classes in the double group Δ

	E	E	$2C_4^2$	$2iC_4$	$2iC_2'$
Δ_1	1	1	1	1	1
Δ_2	1	1	1	−1	−1
Δ_3	1	1	−1	−1	1
Δ_4	1	1	−1	1	−1
Δ_5	2	−2	0	0	0

Figure 146 (p. 678) shows the splitting of the energy levels Γ_8 and Γ_7 upon displacement along the axes Δ, Λ and Σ.

The characters of classes in the double group Λ

	E	E	$2C_3$	$2C_3$	$3iC_2$	$3iC_2$
Λ_1	1	1	1	1	1	1
Λ_2	1	1	1	1	−1	−1
Λ_3	2	2	−1	−1	0	0
Λ_4	1	−1	−1	1	i	$-i$
Λ_5	1	−1	−1	1	$-i$	i
Λ_6	2	−2	1	−1	0	0

The characters of classes in the double group Σ

	E	E	iC_2	iC_2
Σ_1	1	1	1	1
Σ_2	1	1	−1	−1
Σ_3	1	−1	i	$-i$
Σ_4	1	−1	$-i$	i

Table of consistency of representations in the double groups Γ, Δ, Λ and Σ

Γ_6	Γ_7	Γ_8
Δ_5 Λ_6 $\Sigma_3\Sigma_4$	Δ_5 Λ_6 $\Sigma_3\Sigma_4$	$\Delta_5\Delta_5$ $\Lambda_4\Lambda_5\Lambda_6$ $\Sigma_3\Sigma_3\Sigma_4\Sigma_4$

As may be seen from the tables of characters additional representations do not appear on the axes Δ and Σ, while the number of representations on the Λ-axis is doubled (6 instead of 3). The states Γ_6 and Γ_7 along the Δ and Λ axes remain double degenerate, Γ_8 splitting into two double degenerate states. The degeneracy of all the states is completely removed along the Σ axis (Fig. 146).

Let us in conjunction with the double groups turn our attention to the problem of time inversion.

In classical physics the equation of motion is a second-order equation with respect to time

$$\frac{d^2\mathbf{p}}{dt^2}=\mathbf{F}, \qquad (A.208)$$

and for this reason the change of sign of time (the inversion of time) does not change the equation. However, the sign of velocity is changed with the inversion of time. Denote the time inversion operator by $\hat{\tau}$. We may write:

$$\begin{aligned}\hat{\tau}t&=-t;\\ \hat{\tau}\mathbf{v}&=-\mathbf{v}; \qquad (A.209)\\ \hat{\tau}\mathbf{a}&=\mathbf{a}.\end{aligned}$$

Since the wave vector **k** is related to velocity, it follows that $\hat{\tau}\mathbf{k}=-\mathbf{k}$.

Consider the behaviour of energy in case of time inversion. It follows from the time-dependent Schrödinger equation that time inversion is equivalent to the operation of complex conjugation. Indeed,

$$i\hbar=\frac{\partial\psi}{\partial t}=\hat{H}\psi, \qquad (A.210)$$

$$\hat{\tau}i\hbar\frac{\partial\psi}{\partial t}=-i\hbar\frac{\partial\psi}{\partial t}=\bar{H}\psi$$

and

$$-i\hbar\frac{\partial\psi^*}{\partial t}=\hat{H}^*\psi^*. \qquad \text{(A.211)}$$

For a real Hamiltonian $\hat{H}^*=\hat{H}$ the functions ψ and ψ^* satisfy the same equation.

Should the stationary equation

$$\hat{H}\psi=E\psi$$

be written out and the operation of complex conjugation

$$\hat{H}^*\psi^*=E^*\psi^*$$

performed, the solution of the equation for a real Hamiltonian and for real energy values $E^*=E$ would take the form of a double set of wave functions $\{\psi\}$ and $\{\psi^*\}$, each pair corresponding to the same eigenvalue of E. If the sets $\{\psi\}$ and $\{\psi^*\}$ prove different this means that the degree of degeneracy of the level E is not f, but $2f$. If the sets $\{\psi\}$ and $\{\psi^*\}$ "coincide" this means that time inversion does not lead to additional degeneracy.

There may be three different cases of the relationship between the irreducible representations $\Gamma^{(\alpha)}$ and $\Gamma^{(\alpha)*}$ which transform the sets of functions $\{\psi\}$ and $\{\psi^*\}$:

(a) $\Gamma^{(\alpha)}$ and $\Gamma^{(\alpha)*}$ are equivalent, but cannot be real;

(b) $\Gamma^{(\alpha)}$ and $\Gamma^{(\alpha)*}$ are not equivalent;

(c) the irreducible representations $\Gamma^{(\alpha)}$ may be real.

The case (a) points to the existence of a two-fold degeneracy $2f$. In cases (b) and (c) account should be taken of the spin. For an integral spin there is an additional degeneracy in case (b) and none in case (c). For a half-integral spin the picture is reversed—there is additional degeneracy in case (c) and none in case (b).

To check which case should be dealt with one should calculate the sum of characters of the squares of the group elements. The following variants are to be expected:

$$\sum_{G_i}\mathrm{X}(G_i^2)=\begin{cases}-g, & \text{case (a).}\\ 0, & \text{case (b),}\\ g, & \text{case (c).}\end{cases} \qquad \text{(A.212)}$$

RECOMMENDED LITERATURB

1. Anselm, A. I. *Vvedenie v teoriyu poluprovodnikov* (Introduction to the Theory of Semiconductors), Ph. and Math. Publ., Moscow-Leningrad, 1962
2. Bir, G. L., Picus, G. E. *Simmetriya i deformatsionnye effekty v poluprovodnikakh* (The Symmetry and Deformation Effects in Semiconductors), "Nauka", Moscow, 1972
3. Blakemore, J. S. *Semiconductor Statistics*, Pergamon Press, Oxford-London-New York-Paris, 1962
4. Blatt, F. J. *Theory of Mobility of Electrons in Solids*, Academic Press. Inc. Publishers, New York, 1957
5. Callaway, J. *Energy Band Theory*, Academic Press Inc. New York-London, 1964
6. Caplan, J. G. *Simmetriya mnogoelectronnykh sistem* (The Symmetry of Multi-electron Systems), "Nauka", Moscow, 1969
7. Eyring, H., Walter, J. Kimball, G. *Quantum Chemistry*, John Wiley and Sons, Inc., New York-London, 1944
8. Jones, H. *The Theory of Brillouin Zones and Electronic States in Crystals*, North-Holland Publishing Company Amsterdam
9. Kittel, C. *Quantum Theory of Solids*, John Wiley and Sons, Inc., New York-London, 1963
10. Landau, L. D., Lifshitz, E. M., *Quantum Mechanics: Non-Relativistic Theory*, Addison-Wesley, USA, 1965
11. Petrasch, M. I., Trifonov, E. D. *Primenenie teorii grupp v kvantovoi mekhanike* (The Application of the Theory of Groups in Quantum Mechanics), "Nauka", Moscow, 1967
12. Ryvkin, S. M. *Fotoelectricheskie yavleniya v poluprovodnikakh* (Photoelectric Phenomena in Semiconductors), Moscow-Leningrad, 1963
13. Slater, J. C. *Insulators, Semiconductors and Metals*, Mc Graw-Hill Book Company Inc., New York-St Louis-San Francisco-Toronto-London-Sydney, 1967
14. Stilbans, L. S. *Fizika poluprovodnikov* (Semiconductor Physics), "Sov. Radio", Moscow, 1967
15. Streitwolf, H. W. *Gruppentheorie in der Festkörperphysik* (Theory of Groups in Solid State Physics), Akademische Verlagsgesellschaft, Geest und Portig K.-G., Leipzig, 1967
16. Ziman, J. M. *Electrons and Phonons*, Clarendon Press, Oxford, 1960
17. Ziman, J. M. *Principles of the Theory of Solids*, Cambridge, 1964

TO THE READER

Mir Publishers welcome your comments on the content, translation and design of this book.

We would also be pleased to receive any proposals you care to make about our future publications.

Our address:
USSR, 129820, Moscow I-110, GSP
Pervy Rizhsky Pereulok, 2
MIR PUBLISHERS

Printed in the Union of Soviet Socialist Republics

```
{
  "filename": "YV80MDA5NDM3Ny56aXA=",
  "filename_decoded": "a_40094377.zip",
  "filesize": 47280987,
  "md5": "d9d9b24cf45d9f50012406e807abb9ec",
  "header_md5": "c9ff03c276ecbf0d63ef8258724ac7d8",
  "sha1": "b68eae032b4fc63773d1c90b2765c1cac0bbd7fb",
  "sha256": "8679533dc47147275e274222ea6dfdc11f8110fbd8eaea1658f4cd63476bf365",
  "crc32": 1310949965,
  "zip_password": "",
  "uncompressed_size": 47985195,
  "pdg_dir_name": "a_40094377",
  "pdg_main_pages_found": 694,
  "pdg_main_pages_max": 694,
  "total_pages": 695,
  "total_pixels": 2649584640,
  "pdf_generation_missing_pages": false
}
```